

Scalable Bayes via Barycenter in Wasserstein Space

Sanvesh Srivastava ^{*1}, Cheng Li ^{†2}, and David B. Dunson ^{‡3}

¹*Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa, USA*

²*Department of Statistics and Applied Probability, National University of Singapore, Singapore*

³*Department of Statistical Science, Duke University, Durham, North Carolina, USA*

April 27, 2017

Abstract

Divide-and-conquer based methods for Bayesian inference provide a general approach for tractable posterior inference when the sample size is large. These methods first divide the data into smaller subsets and sample from the posterior distribution of parameters in parallel across all subsets, and then combine posterior samples from all the subsets to approximate the full data posterior distribution. Sampling in the first step is more efficient than sampling from the full data posterior due to smaller size of any subset. Since the combination step takes negligible time relative to sampling, posterior computations can be scaled to massive data by dividing the full data into sufficiently large number of data subsets. One such approach relies on the geometry of posterior distributions estimated across different subsets and combines them through their barycenter in a Wasserstein space of probability measures. We provide theoretical guarantees on the accuracy of approximation that are applicable in many applications. We show that the geometric method approximates the full data posterior distribution better than its competitors across diverse simulations and reproduces known results when applied to a movie ratings database.

Key words: barycenter; big data; distributed Bayesian computations; empirical measures; linear programming; optimal transportation; Wasserstein distance; Wasserstein space.

1 Introduction

Developing efficient sampling algorithms is an active area of research motivated by tractable Bayesian inference in large sample settings. Sampling remains a primary tool for inference in Bayesian models, with Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) providing two broad classes of algorithms that are routinely used. Most MCMC and SMC algorithms face problems in scaling up to massive data settings due to memory and computational bottlenecks that arise; this has motivated a rich literature in recent years proposing a variety of strategies to enable better performance in such settings. Our focus is on proposing a very general divide-and-conquer technique, which is designed to combine results from any posterior sampling algorithm applied in parallel using subsets of the data.

Massive data pose three major problems for existing sampling algorithms. First, if full data require multiple machines for storage, then any sampler can access only a small fraction of the data and posterior sampling given the full data is expensive due to extensive communication among machines. Second, if full

^{*}sanvesh-srivastava@uiowa.edu

[†]stalic@nus.edu.sg

[‡]dunson@duke.edu

data are available to the sampler, then sampling is infeasible because computation of Hessians and acceptance ratios often scale as $O(n^3)$, where n is the sample size. Finally, sampling in hierarchical Bayesian models requires generation of $O(n)$ latent variables, which becomes inefficient as n increases. A variety of methods exist to address these issues using optimization and sampling.

Optimization-based methods for Bayesian inference obtain an analytic approximation of the full data posterior distribution. The two most common techniques are polynomial approximation (Rue et al., 2009) and projection of the full data posterior distribution on a class of distributions with analytically tractable posterior densities, which includes variational Bayes and expectation propagation (Wainwright and Jordan, 2008; Gelman et al., 2014). Both techniques estimate parameters of the approximate distribution using a variety of optimization algorithms (Tan and Nott, 2013; Wand, 2014). Stochastic approximation significantly improves the efficiency of estimation by accessing the data in small batches and updating the parameter estimates sequentially (Broderick et al., 2013; Hoffman et al., 2013); however, optimization can be nontrivial for complex likelihoods frequently used in hierarchical models. Further, variational Bayes and expectation propagation often have excellent predictive performance but can be highly biased in estimation of posterior uncertainty and dependence.

There is extensive work in sampling-based methods for Bayesian inference. The three main techniques used in these methods are as follows. First, subsampling-based methods obtain posterior samples conditioned on a small fraction of the data (Maclaurin and Adams, 2015). Coupling of subsampling with modified Hamiltonian or Langevin Dynamics improves posterior exploration and convergence to the stationary distribution (Welling and Teh, 2011; Ahn et al., 2012; Korattikara et al., 2014; Lan et al., 2014; Shahbaba et al., 2014); see Bardenet et al. (2015) for a review. Second, the exact transition kernel in posterior sampling is replaced by an approximation that significantly reduces the time required to finish an iteration of the sampler (Johndrow et al., 2015; Alquier et al., 2016). Finally, divide-and-conquer approaches first divide the data into smaller subsets and sample in parallel across subsets, and then combine the posterior samples from all the subsets. Our focus is on scalable Bayesian methods based on the divide-and-conquer technique. These methods have two subgroups that differ mainly in their sampling scheme for every subset and their method for combining posterior samples obtained from all the subsets.

The first subgroup modifies the prior to sample from the posterior distribution of the parameter conditioned on a data subset. Let k be the number of subsets, $\pi(\theta)$ be the prior density of parameter θ , and $l_i(\theta)$ be the likelihood for subset i ($i = 1, \dots, k$). Samples from subset posterior distribution i are obtained using $l_i(\theta)$ and $\pi(\theta)^{1/k}$ as the likelihood and prior. Consensus Monte Carlo combines subset posterior samples by averaging (Scott et al., 2016). This relies heavily on the normality assumption, which is relaxed using a combination based on kernel density estimation (Neiswanger et al., 2014). Both methods perform poorly if the supports of subset posteriors are different, which motivates the combination using the Weierstrass transform and random partition trees (Wang and Dunson, 2013; Wang et al., 2015). These methods offer simple approaches to combine samples from subset posterior distributions but have two major limitations. First, the sampling algorithm depends on the model parameterization. Second, transforming $\pi(\theta)$ to $\pi(\theta)^{1/k}$ makes posterior computations harder because existing off-the-shelf sampling algorithms cannot be used directly.

The second subgroup modifies the subset likelihood to sample from a subset posterior distribution and combines samples from subset posterior distributions through their geometric center. These methods modify the likelihood to $l_i(\theta)^k$ and use prior $\pi(\theta)$ to sample from subset posterior distribution i ($i = 1, \dots, k$). M-Posterior combines subset posterior distributions through their median in the Wasserstein space of order 1 (Minsker et al., 2014). The robustness of the median implies that it could ignore valuable information in some subset posterior distributions, which motivates combination through the mean in the Wasserstein space of order 2 called Wasserstein Posterior (WASP) (Srivastava et al., 2015). The WASP approach strikes a balance between the generality of sampling and the efficiency of optimization; however, its computations are developed for independent identically distributed (*iid*) data and its theoretical properties are unknown.

Our main goal is to study three theoretical properties of WASP and apply WASP in a variety of practical

problems. The *iid* assumption of WASP rules out many important practical problems, including regression and classification, where the data are independent and non-identically distributed (*inid*). We relax this assumption and our results are applicable to any *inid* data. Second, we show that if the number of subsets are chosen appropriately, then the WASP achieves almost the same rate of convergence as that of the full data posterior distribution. This implies that WASP can be used as an efficient alternative to the full data posterior distribution for uncertainty quantification in massive data settings. Third, we show that the method for estimating WASP is independent of the form of the model, so WASP is very general and can be easily used for estimating posterior summaries for any functional of the model parameters. We emphasize that WASP is not a new sampling algorithm but a general approach to easily extend existing sampling algorithms for massive data applications.

2 Preliminaries

2.1 Wasserstein space, Wasserstein distance, and Wasserstein barycenter

We recall elementary properties and definitions related to the Wasserstein space of probability measures. Let (Θ, ρ) be a complete separable metric space and $\mathcal{P}(\Theta)$ be the space of all probability measures on Θ . The Wasserstein space of order 2 is defined as

$$\mathcal{P}_2(\Theta) := \left\{ \mu \in \mathcal{P}(\Theta) : \int_{\Theta} \rho(\theta_0, \theta)^2 \mu(d\theta) < \infty \right\}, \quad (1)$$

where $\theta_0 \in \Theta$ is arbitrary and $\mathcal{P}_2(\Theta)$ does not depend on the choice of θ_0 . The space $\mathcal{P}_2(\Theta)$ is equipped with a natural distance between its elements. Let $\mu, \nu \in \mathcal{P}_2(\Theta)$ and $\Pi(\mu, \nu)$ be the set of all probability measures on $\Theta \times \Theta$ with marginals μ and ν , then the Wasserstein distance of order 2 between μ and ν is defined as

$$W_2(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\Theta \times \Theta} d^2(x, y) d\pi(x, y) \right)^{\frac{1}{2}}. \quad (2)$$

In our applications ρ is the Euclidean metric and we refer to $\mathcal{P}_2(\Theta)$ and W_2 as the Wasserstein space and the Wasserstein distance without explicitly mentioning their order. If Π_1, \dots, Π_k are a collection of probability measures in $\mathcal{P}_2(\Theta)$, then their barycenter in $\mathcal{P}_2(\Theta)$ is defined as

$$\bar{\Pi} = \operatorname{argmin}_{\Pi \in \mathcal{P}_2(\Theta)} \sum_{j=1}^k \frac{1}{k} W_2^2(\Pi, \Pi_j). \quad (3)$$

This generalizes the Euclidean barycenter, which is the sample mean, to $\mathcal{P}_2(\Theta)$ (Agueh and Carlier, 2011). The barycenter $\bar{\Pi}$ is analytically intractable, except in few special cases. Let $\delta_a(x) = 1$ if $a = x$ and 0 otherwise. If X_{j1}, \dots, X_{jm} are samples from Π_j ($j = 1, \dots, k$), then $\hat{\Pi}_j(\cdot) = \sum_{i=1}^m \delta_{X_{ji}}(\cdot)/m$ is an empirical measure that approximates Π_j ($j = 1, \dots, k$). If $\bar{\Pi}$ is assumed to be an empirical measure, then the optimization problem in (3) reduces to a linear program; see Cuturi and Doucet (2014), Carlier et al. (2015), and Srivastava et al. (2015) for different algorithms to solve this linear program.

2.2 Stochastic approximation and subset posterior density

Consider a general set-up for *inid* data. Let $Y^{(n)} = (Y_1, \dots, Y_n)$ be n observations and the distribution of Y_i is $P_{\theta, i}$, $i = 1, \dots, n$, where θ lies in the parameter space $\Theta \subset \mathbb{R}^p$. Assume that $P_{\theta, i}$ has density $p_i(\cdot|\theta)$ with respect to the Lebesgue measure, so $dP_{\theta, i}(y_i) = p_i(y_i|\theta)dy_i$ and the likelihood given $Y^{(n)}$

is $l(\theta) = \prod_{i=1}^n p_i(y_i|\theta)$. Given a prior distribution Π on Θ that has density π with respect to the Lebesgue measure, the posterior density of θ given $Y^{(n)}$ using Bayes theorem is

$$\pi(\theta | Y^{(n)}) = \frac{\prod_{i=1}^n p_i(y_i | \theta)\pi(\theta)}{\int_{\Theta} \prod_{i=1}^n p_i(y_i | \theta)\pi(\theta)d\theta} = \frac{l(\theta)\pi(\theta)}{\int_{\Theta} l(\theta)\pi(\theta)d\theta}. \quad (4)$$

In most cases $\pi(\theta | Y^{(n)})$ is analytically intractable, and accurate approximations of $\pi(\theta | Y^{(n)})$ are obtained using Monte Carlo methods, such as importance sampling and MCMC, and deterministic approximations, such as Laplace's method and variational Bayes. For example, in the context of logistic regression, $P_{\theta,i}$ is the Bernoulli distribution with mean $1/\{1 + \exp(-x_i^T \theta)\}$, where x_i^T is the i th row of the design matrix $X \in \mathbb{R}^{n \times p}$ and $\Theta = \mathbb{R}^p$. The posterior density of θ is analytically intractable, and it is typical to rely on Gibbs samplers based on data augmentation. These samplers introduce latent variables $\{z_i, i = 1, \dots, n\}$ and alternately sample the latent variables and the parameters from their full conditional distributions. Related algorithms are very common and are computationally prohibitive for large n because they require repeated passes through the whole data.

Divide-and-conquer-type methods resolve this problem by partitioning the data into smaller subsets. Let k be the number of subsets. The default strategy is to randomly allocate samples to subsets. Let $Y_{[j]} \equiv Y_j^{(m_j)} = (Y_{j1}, \dots, Y_{jm_j})$ denote data on the j th subset, where m_j is the size of the j th subset and $\sum_{j=1}^k m_j = n$. We assume that $m_j = m$ ($j = 1, \dots, k$) for ease of presentation, so $n = km$, the likelihood given $Y_{[j]}$ is $l_j(\theta) = \prod_{i=1}^m p_{ji}(y_{ji}|\theta)$, and $l(\theta)$ in (4) equals $\prod_{j=1}^k l_j(\theta)$. Minsker et al. (2014) and Srivastava et al. (2015) define subset posterior density j given $Y_{[j]}$ as

$$\pi_m(\theta | Y_{[j]}) = \frac{\{\prod_{i=1}^m p_{ji}(y_{ji}|\theta)\}^\gamma \pi(\theta)}{\int_{\Theta} \{\prod_{i=1}^m p_{ji}(y_{ji}|\theta)\}^\gamma \pi(\theta)d\theta} = \frac{l_j(\theta)^\gamma \pi(\theta)}{\int_{\Theta} l_j(\theta)^\gamma \pi(\theta)d\theta}, \quad (5)$$

where γ is a positive real number such that $g_1 \gamma m \leq n \leq g_2 \gamma m$ for some $g_1, g_2 > 0$. In the present context, we assume that $\gamma = k$ with $g_1 = g_2 = 1$ following Minsker et al. (2014); more general conditions on γ are defined later in Section 3.2. This modified form of subset posterior compensates for the fact that j th subset has access to only (m/n) -fraction of the full data and ensures that $\pi_m(\theta | Y_{[j]})$ and $\pi_n(\theta | Y^{(n)})$ in (4) have variances of the same order. Minsker et al. (2014) refer to this as *stochastic approximation* because raising $l_j(\theta)$ ($j = 1, \dots, k$) to the power γ is equivalent to replicating every X_{ji} ($i = 1, \dots, m$) γ -times so that $\pi_m(\theta | Y_{[j]})$ ($j = 1, \dots, k$) are noisy approximations of $\pi(\theta | Y^{(n)})$.

One advantage of using stochastic approximation to define $\pi_m(\theta | Y_{[j]})$ in (5) is that off-the-shelf sampling algorithms can be used directly even when the prior density is the form of a discrete mixture. Consider a simple example of univariate density estimation using Dirichlet process mixtures of Gaussians. Let X_i ($i = 1, \dots, n$) be *iid* samples from a distribution P_0 with density p_0 . The data are randomly split into k subsets of equal size m . The truncated stick-breaking representation of Dirichlet process prior implies that the prior distribution Π on \mathcal{P} has a finite mixture representation, where \mathcal{P} is the set of probability distributions that have a density. The subset posterior density j of our competitors is proportional to $l_j(p)\pi(p)^{1/k}$, where p is the density of a distribution $P \in \mathcal{P}$, $l_j(p)$ is the likelihood for subset j , and $\pi(p)$ has the form of a Dirichlet process mixture. In this case, existing Gibbs sampling algorithms cannot be applied directly. In contrast, we show in Appendix C that modification of the likelihood using stochastic approximation in WASP allows the use of existing Gibbs sampling algorithm for density estimation.

Stochastic approximation does not add any extra burden to the computations required for sampling from the subset posterior distribution of θ conditioned on m observations. We raise the likelihood in every subset to the power γ . This is equivalent to replicating observations γ -times, which seems to offset the benefits of partitioning. However, the replication of observation is not required in implementation of the sampler; we simply modify the likelihood in the full data sampler by raising it to the power γ . For example,

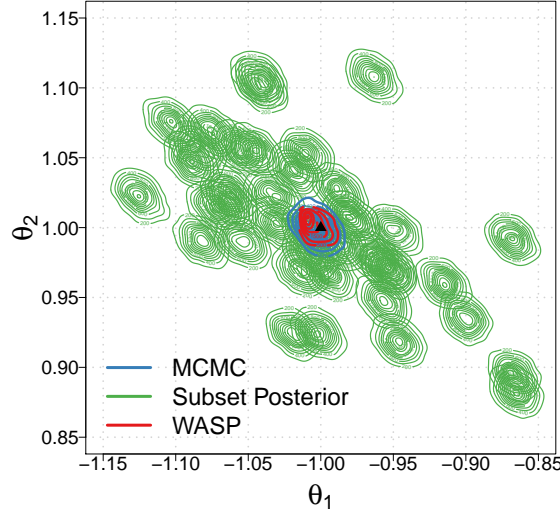


Figure 1: Binned kernel density estimates of full data posterior distribution, subset posterior distributions, and WASP for coefficients (θ_1, θ_2) in logistic regression. The x and y axes represent posterior samples for θ_1 and θ_2 . The true values of θ_1 and θ_2 are -1 and 1 (black triangle).

stochastic approximation is easily implemented using the `increment_log_prob` function in Stan (Stan Development Team, 2014). We provide more examples for a variety of models in Section 4.

A simple logistic regression example demonstrates that $\pi_m(\theta \mid Y_{[j]})$ in (5) is a noisy approximation of $\pi(\theta \mid Y^{(n)})$ in (4). We simulated data for logistic regression with $n = 10^5$, $p = 2$, $\theta = (-1, 1)^T$, and entries of X randomly set to ± 1 (Figure 1). We set $\gamma = k = 40$ and obtained samples of θ from $\pi(\theta \mid Y^{(n)})$ and from $\pi_m(\theta \mid Y_{[j]})$ ($j = 1, \dots, k$) using the Stan’s HMC sampling algorithm. The contours for the subset and full data posterior densities are very similar, indicating all densities have similar spreads. We also notice that subset posteriors are noisy approximations of the full data posterior in that most of them have a bias and do not concentrate at the true θ .

3 Wasserstein Posterior (WASP): The general framework

3.1 Definition and estimation of the WASP

The WASP approach combines subset posterior distributions $\Pi_m(\cdot \mid Y_{[j]})$ ($j = 1, \dots, k$) through their barycenter in $\mathcal{P}_2(\Theta)$, where the density of $\Pi_m(\cdot \mid Y_{[j]})$ is $\pi_m(\cdot \mid Y_{[j]})$ in (5). The barycenter represents a geometric center of a collection of probability distributions that can be efficiently computed using a linear program. Motivated by this, Srivastava et al. (2015) proposed to combine a collection of subset posterior distributions through their barycenter in the Wasserstein space called *WASP*. Assuming that subset posterior distributions $\Pi_m(\cdot \mid Y_{[j]})$ ($j = 1, \dots, k$) have finite second moments, the WASP is defined using (3) as

$$\bar{\Pi}_n(\cdot \mid Y^{(n)}) = \underset{\Pi \in \mathcal{P}_2(\Theta)}{\operatorname{argmin}} \sum_{j=1}^k \frac{1}{k} W_2^2\{\Pi, \Pi_m(\cdot \mid Y_{[j]})\}. \quad (6)$$

The WASP is analytically tractable only in special cases, but it can be estimated using a linear program if the subset posterior distributions have an atomic form (Srivastava et al., 2015). Let $\{\theta_{j1}, \dots, \theta_{js}\}$ be

the θ samples obtained from subset posterior density j in (6) using a sampling algorithm, including HMC, MCMC, SMC, or importance sampling. Approximate j th subset posterior distribution $\Pi_m(\cdot \mid Y_{[j]})$ using the empirical measure

$$\hat{\Pi}_m(\cdot \mid Y_{[j]}) = \sum_{i=1}^S \frac{1}{S} \delta_{\theta_{ji}}(\cdot) \quad (j = 1, \dots, k). \quad (7)$$

Srivastava et al. (2015) approximate the WASP as

$$\hat{\Pi}_n(\cdot \mid Y^{(n)}) = \sum_{j=1}^k \sum_{i=1}^S a_{ji} \delta_{\theta_{ji}}(\cdot), \quad 0 \leq a_{ji} \leq 1, \quad \sum_{j=1}^k \sum_{i=1}^S a_{ji} = 1, \quad (8)$$

where a_{ji} ($j = 1, \dots, k$; $i = 1, \dots, S$) are unknown weights of the atoms. There are many specialized algorithms to estimate the WASP that exploit the structure of the linear program in (6) when $\Pi_m(\cdot \mid Y_{[j]})$ and $\hat{\Pi}_n(\cdot \mid Y^{(n)})$ are restricted to have atomic forms in (7) and (8), respectively; for example, Cuturi and Doucet (2014) extend the Sinkhorn algorithm of Cuturi (2013) using entropy-smoothed sub-gradient methods, Carlier et al. (2015) develop a non-smooth optimization algorithm, and Srivastava et al. (2015) propose an efficient linear program that exploits the sparsity of constraints to solve (6).

3.2 Theoretical properties of the WASP

We study asymptotic properties of the WASP defined in (6) that are important for applications in scalable Bayesian inference. First, the rate of concentration of the WASP around the true value of the parameter is unknown. Let θ_0 be the true parameter, then the full data posterior distribution $\Pi_n(\cdot \mid Y^{(n)})$ converges to δ_{θ_0} in Hellinger distance at the optimal parametric rate of $n^{-1/2}$ up to logarithm factors (Ghosal et al., 2000); however, similar guarantees are unknown for the WASP. In addition, in many applications, the parameter θ is not of immediate interest and the interest lies in $f(\theta)$ for some known function f . We do not know if the WASP of k subset posterior distributions for $f(\theta)$ converges to $\delta_{f(\theta_0)}$. The current asymptotic justification of WASP relies only on posterior consistency under an *iid* assumption; for example, Theorem 3.3 in Srivastava et al. (2015). We obtain much stronger asymptotic support by showing rates in an *inid* case, including for functionals of the original parameters.

Our theoretical setup is based on the following assumptions.

(A1) Θ is a compact space in ρ metric, and θ_0 is an interior point of Θ ; $g_1 \gamma m \leq n \leq g_2 \gamma m$ for some constants $g_1, g_2 > 0$.

(A2) For any $\theta, \theta' \in \Theta$ and $j = 1, \dots, m$, there exist positive constants α and C_L , such that

$$h_{mj}^2(\theta, \theta') \geq C_L \rho^{2\alpha}(\theta, \theta'),$$

where $h_{mj}^2(\theta, \theta')$ is the pseudo Hellinger distance defined in A.1.

(A3) (Entropy Condition) There exist constants $D_1 > 0$, $0 < D_2 < D_1^2/2^{12}$, a function $\Psi(u, r) \geq 0$ that is nonincreasing in $u \in \mathbb{R}^+$ and nondecreasing in $r \in \mathbb{R}^+$, such that for all $j = 1, \dots, k$, for any $u, r > 0$ and for all sufficiently large m ,

$$H_{\square}(u, \{\mathbf{p}_j(\mathbf{y}|\theta) : \theta \in \Theta, h_{mj}(\theta, \theta_0) \leq r\}, h_{mj}) \leq \Psi(u, r) \text{ for all } j = 1, \dots, k;$$

$$\text{and } \int_{D_1 r^2/2^{12}}^{D_1 r} \sqrt{\Psi(u, r)} du < D_2 \sqrt{mr^2},$$

where $\mathbf{p}_j(\mathbf{y}|\theta) = \{p_{j1}(y_{j1} \mid \theta), \dots, p_{jm}(y_{jm} \mid \theta)\}^T$ and H_{\square} is the h_{mj} -bracketing entropy of the set $\{\mathbf{p}_j(\mathbf{y}|\theta) : \theta \in \Theta, h_{mj}(\theta, \theta_0) \leq r\}$, which is defined in A.2.

(A4) (Prior Thickness) There exist positive constants κ and c_π , such that uniformly over all $j = 1, \dots, k$,

$$\Pi \left(\theta \in \Theta : \frac{1}{m} \sum_{i=1}^m E_{P_{\theta_0}} \exp \left(\kappa \log_+ \frac{p_{ji}(Y_{ji}|\theta_0)}{p_{ji}(Y_{ji}|\theta)} \right) - 1 \leq \frac{\log^2 m}{m} \right) \geq \exp(-c_\pi k \log^2 m)$$

where $\log_+ x = \max(\log x, 0)$ for $x > 0$.

(A5) The metric ρ satisfies that for any $N \in \mathbb{N}$, $\theta_1, \dots, \theta_N, \theta' \in \Theta$ and nonnegative weights $\sum_{i=1}^N w_i = 1$,

$$\rho \left(\sum_{i=1}^N w_i \theta_i, \theta' \right) \leq \sum_{i=1}^N w_i \rho(\theta_i, \theta').$$

For theory development, we have assumed compact support in (A1) and lower bounded pseudo Hellinger distance in (A2), which is similar to Theorem 10 in Ghosal and van Der Vaart (2007). Typically, $\alpha = 1$ for most regular models; e.g., generalized linear models. If the model is non-regular, then α can be less than 1. For example, the densities may have discontinuities depending on the parameter (see Ibragimov and Has'minskii (1981) Chapter V and VI). (A3) parallels the entropy condition used in Theorem 1 of Wong and Shen (1995), which has been adapted here for the *inid* setup using the generalized bracketing entropy, and will simplify to a similar entropy condition to that in Theorem 1 of Wong and Shen (1995) if the data are *iid*. The convexity property of ρ in (A5) is mainly used to establish an averaging inequality under W_2 distance and is satisfied by for example, the Euclidean metric and ℓ_q metric with $q \geq 1$; see Lemma 1.7 in the Supplementary Material.

The two theorems below describe two asymptotic properties of WASP for models that satisfy assumptions (A1)–(A4) above. The asymptotic behavior of the WASP for *inid* data is described in two steps. The first theorem describes the asymptotic behavior of subset posterior distributions $\Pi_m(\cdot | Y_{[j]})$ ($j = 1, \dots, k$). The second theorem describes the asymptotic behavior of the WASP using the asymptotic behavior of subset posterior distributions.

Theorem 3.1 *If Assumptions (A1)–(A4) hold for the j th subset posterior $\Pi_m(\cdot | Y_{[j]})$ with $j = 1, \dots, k$, then there exists constants α and C_1 that do not depend on j , such that as $m \rightarrow \infty$,*

$$E \left[W_2^2 \{ \Pi_m(\cdot | Y_{[j]}), \delta_{\theta_0}(\cdot) \} \right] \leq C_1 \left(\frac{\log^2 m}{m} \right)^{\frac{1}{\alpha}}, \quad (9)$$

where expectation is with respect to probability measure $P_{\theta_0}^{(m)}$.

Theorem 3.1 proves posterior convergence in expectation $E_{P_{\theta_0}^{(m)}}$, which is stronger than the commonly studied posterior convergence in probability. Assumption (A4) is crucial in providing a stronger control over the tail probability as the posterior probability mass moves away from the true parameter θ_0 , typically with an exponentially decaying rate. For regular models with $\alpha = 1$, Theorem 3.1 says that the W_2 distance between the subset posterior $\Pi_m(\cdot | Y_{[j]})$ and the delta measure at θ_0 shrinks to zero at a nearly parametric rate $m^{-1/2}$ up to some logarithm factors. The $m^{-1/2}$ rate is the optimal parametric rate because each subset only has sample size m . For nonregular models with $\alpha < 1$, the convergence rate can be faster than parametric, which achieves the same rate as the frequentist maximum likelihood estimator up to some logarithm factors; see Ghosal et al. (1995), Ibragimov and Has'minskii (1981). A combination of k disjoint subsets yields the following convergence result for the WASP.

Theorem 3.2 *If Assumptions (A1)-(A5) hold for all subset posteriors $\Pi_m(\cdot \mid Y_{[j]})$ with $j = 1, \dots, k$, then as $m \rightarrow \infty$,*

$$W_2 \left\{ \bar{\Pi}_n(\cdot \mid Y^{(n)}), \delta_{\theta_0}(\cdot) \right\} = O_p \left(\sqrt{\frac{\log^{2/\alpha} m}{m^{1/\alpha}}} \right), \quad (10)$$

where O_p is in $P_{\theta_0}^{(n)}$ -probability.

Theorem 3.2 shows that the WASP of k subset posterior distributions converges to the true parameter θ_0 at a rate of $m^{-\frac{1}{2\alpha}}$ up to logarithm factors. When $\alpha = 1$, as in regular models, this rate is about the same as the rate obtained in Corollary 3.12 and Theorem 3.14 of Minsker et al. (2014) for the M-Posterior. If $\alpha = 1$ and the number of subsets k increases slowly with n , for example $k = O(\log^c n)$ for some constant $c > 0$, then Theorem 3.2 implies that the WASP has a near optimal convergence rate $O_p \left(\sqrt{\frac{\log^{c+2} n}{n}} \right)$, since the optimal parametric rate on the full dataset is $O_p(n^{-1/2})$.

Suppose $f : \Theta \mapsto \mathbb{R}^q$ is a function that maps θ to $\{f_1(\theta), \dots, f_q(\theta)\}$, where $q \geq 1$ is a positive integer. A direct application of Lemma 8.5 in Bickel and Freedman (1981) gives the following corollary about the WASP of a function of θ . As long as the function is bounded almost linearly by the ρ metric in (1), its WASP possesses the same posterior convergence rate as in Theorem 3.2.

Corollary 3.3 *Suppose $f(\cdot) = \{f_1(\cdot), \dots, f_q(\cdot)\}$ is a function that maps $\Theta \mapsto \mathbb{R}^q$ such that $|f(\theta)|^2 = \sum_{i=1}^q \{f_i(\theta)\}^2 \leq C_f \{1 + \rho^2(\theta, \theta_0)\}$, where $C_f > 0$ is a fixed constant. If the conditions in Theorem 3.2 hold and $\bar{f}_{\#} \bar{\Pi}_n(\cdot \mid Y^{(n)})$ represents the WASP of the subset posterior distributions for $f(\theta)$, then as $m \rightarrow \infty$,*

$$W_2 \left\{ \bar{f}_{\#} \bar{\Pi}_n(\cdot \mid Y^{(n)}), \delta_{f(\theta_0)}(\cdot) \right\} = O_p \left(\sqrt{\frac{\log^{2/\alpha} m}{m^{1/\alpha}}} \right).$$

3.3 Computation of the WASP

Corollary 3.3 is very useful in applications because it says that the combination step in the WASP is independent of the model parametrization. Let $f_{\#} \Pi_m(\cdot \mid Y_{[j]})$ be the j th subset posterior distribution for $f(\theta)$ ($j = 1, \dots, k$), then the WASP of k subset posterior distributions converges to $f(\theta_0)$ at the rate obtained in Theorem 3.2. In practice, we have S_j posterior samples of θ obtained from subset posterior j denoted as θ_{ji} ($i = 1, \dots, s_j; j = 1, \dots, k$). Algorithm 1 estimates an atomic approximation of $\bar{f}_{\#} \bar{\Pi}_n(\cdot \mid Y^{(n)})$, denoted as $\hat{\bar{f}}_{\#} \hat{\bar{\Pi}}_n(\cdot \mid Y^{(n)})$, based on the subset posterior samples $f(\theta_{ji})$ ($i = 1, \dots, s_j; j = 1, \dots, k$). The atomic form of the WASP is supported on a grid with mesh-size ϵ estimated from the subset posterior samples of $f(\theta)$. Algorithm 1 estimates the weights of the atoms located on the grid by solving a discrete version of (6). The support of $\hat{\bar{f}}_{\#} \hat{\bar{\Pi}}_n(\cdot \mid Y^{(n)})$ scales exponentially in k , but theoretical properties of discrete barycenters imply that $\hat{\bar{f}}_{\#} \hat{\bar{\Pi}}_n(\cdot \mid Y^{(n)})$ is supported only on $O(k)$ elements of the grid; see Theorem 2 in Anderes et al. (2016). We exploit this sparsity by adapting the algorithm in Srivastava et al. (2015) and by relying on Gurobi to exploit sparsity in linear programs (Gurobi Optimization Inc., 2014). A general algorithm that exploits the sparsity in the support of the atomic approximation of WASP is left for future research. In all our experiments, $\hat{\bar{f}}_{\#} \hat{\bar{\Pi}}_n(\cdot \mid Y^{(n)})$ provides an approximation of the full data posterior distribution of $f(\theta)$.

Algorithm 1 Estimation of the WASP for $f(\theta)$ in Corollary 3.3 given samples of θ from k subset posteriors

Input: Samples from k subset posteriors, $\{\theta_{ji} : \theta_{ji} \sim \Pi_m(\cdot | Y_{[j]})$, $i = 1, \dots, s_j$, $j = 1, \dots, k\}$; mesh size $\epsilon > 0$.

Do:

1. Define $\phi_i^j = (\phi_{i1}^j, \dots, \phi_{iq}^j) = f(\theta_{ji})$ ($i = 1, \dots, s_j$; $j = 1, \dots, k$), the matrix of atoms of subset posterior j , $\Phi_j \in \mathbb{R}^{s_j \times q}$, with ϕ_i^j as row i ($i = 1, \dots, s_j$). For $r = 1, \dots, q$, let $\phi_{\min} = (\phi_{\min 1}, \dots, \phi_{\min q})$ with $\phi_{\min r} = \min_{ji} \phi_{ir}^j$, and $\phi_{\max} = (\phi_{\max 1}, \dots, \phi_{\max q})$ with $\phi_{\max r} = \max_{ji} \phi_{ir}^j$.
2. Set the number of atoms in the empirical approximation for the WASP $g = g_1 \times \dots \times g_q$, where $g_r = \lceil \frac{\phi_{\max r} - \phi_{\min r}}{\epsilon} \rceil$ ($r = 1, \dots, q$).
3. Define the matrix of WASP atoms $\overline{\Phi} \in \mathbb{R}^{g \times q}$ with rows formed by stacking vectors

$$\left\{ \phi_{\min 1} + \frac{i_1}{g_1} (\phi_{\max 1} - \phi_{\min 1}), \dots, \phi_{\min q} + \frac{i_q}{g_q} (\phi_{\max q} - \phi_{\min q}) \right\}, \quad (i_r = 1, \dots, g_r; r = 1, \dots, q).$$

4. Set the distance matrix between the atoms of WASP and the j th subset posterior, $D_j \in \mathbb{R}_+^{g \times s_j}$, as

$$(D_j)_{uv} = \sum_{r=1}^q (\overline{\Phi}_{ur} - \phi_{vr}^j)^2, \quad (u = 1, \dots, g; v = 1, \dots, s_j; j = 1, \dots, k),$$

where $\overline{\Phi}_{ur}$ is the (u, r) -entry of $\overline{\Phi}$.

5. Estimate $\hat{\alpha}_1, \dots, \hat{\alpha}_g$ by solving the linear program (31) in Appendix D.

Return: $\hat{f}_{\#}^{\Pi}(\cdot | Y^{(n)}) = \sum_{i=1}^g \hat{\alpha}_i \delta_{\overline{\Phi}_i}(\cdot)$, the atomic approximation of $\overline{f}_{\#}^{\Pi_n}(\cdot | Y^{(n)})$.

4 Experiments

4.1 Setup

We compared WASP with consensus Monte Carlo (CMC) (Scott et al., 2016), semiparametric density product (SDP) (Neiswanger et al., 2014), and variational Bayes (VB). The sample sizes and the number of parameters in our experiments were chosen such that sampling from the full data posterior distribution was computationally feasible. Every sampling algorithm was run 10,000 iterations. We discarded the first 5,000 samples as burn-in and thinned the chain by collecting every fifth sample. Convergence of the chains to their stationary distributions was confirmed using trace plots. All experiments were run on an Oracle Grid Engine cluster with 2.6GHz 16 core compute nodes. Full data posterior computations were allotted memory resources of 64GB, and all other methods were allotted memory resources of 16GB.

The sampling algorithm for the full data posterior was modified to obtain samples from the subset posteriors in CMC, SDP, and WASP. The sampling algorithms for subset posteriors in CMC and SDP were the same and were based on Equation (2) in Scott et al. (2016). The sampling algorithm for subset posteriors in WASP was based on (5). Samples from the approximate posterior distributions of θ in CMC, SDP, and WASP were obtained in two steps. First, samples from subset posteriors of θ were obtained in parallel across k subsets. Second, the samples of θ from all the subsets were combined using implementations of CMC and SDP in `parallelMCMC` package (Miroshnikov and Conlon, 2014) and using Algorithm 1 for the WASP.

The full data posterior distribution obtained using MCMC served as the benchmark in all our comparisons. Let $\pi(\theta | Y^{(n)})$ be the density of the full data posterior distribution for θ estimated using sampling and $\hat{\pi}(\theta | Y^{(n)})$ be the density of an approximate posterior distribution for θ estimated using the WASP or its competitors. We used the following metric based on the total variation distance to compare the accuracy $\hat{\pi}(\theta | Y^{(n)})$ in approximating $\pi(\theta | Y^{(n)})$

$$\text{accuracy} \left\{ \hat{\pi}(\theta | Y^{(n)}) \right\} = 1 - \frac{1}{2} \int_{\Theta} \left| \hat{\pi}(\theta | Y^{(n)}) - \pi(\theta | Y^{(n)}) \right| d\theta. \quad (11)$$

The accuracy metric lies in $[0, 1]$ (Faes et al., 2012). The approximation of full data posterior density by $\hat{\pi}$ is poor or excellent if the accuracy metric is close to 0 or 1, respectively. In our experiments, we computed the kernel density estimates of $\hat{\pi}$ and π from the posterior samples of θ using R package `KernSmooth` (Wand, 2015) and calculated the integral in (11) using numerical approximation.

4.2 Simulated data: finite mixture of Gaussians

Finite mixture of Gaussians are widely used for model-based classification, clustering, and density estimation (Fraley and Raftery, 2002). Let n , p , and L be the sample size, the dimension of observations, and the number of mixture components. If $\mathbf{y}_i \in \mathbb{R}^p$ is the i th observation ($i = 1, \dots, n$), then the mixture of L Gaussians assumes that any $\mathbf{y} \in \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ is generated from the density

$$f_{\text{mix}}(\mathbf{y} | \theta) = \sum_{l=1}^L \pi_l \mathcal{N}_p(\mathbf{y} | \boldsymbol{\mu}_l, \Sigma_l), \quad (12)$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_L)$ lies in a $(L - 1)$ -simplex, $\boldsymbol{\mu}_l$ and Σ_l ($l = 1, \dots, L$) are the mean and covariance parameters of a p -variate Gaussian distribution, and $\theta = \{\boldsymbol{\pi}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_L, \Sigma_1, \dots, \Sigma_L\}$. We set $L = 2$ and $p = 2$ and simulated data from (43) using $\boldsymbol{\pi} = (0.3, 0.7)$, $\boldsymbol{\mu}_1 = (1, 2)^T$, $\boldsymbol{\mu}_2 = (7, 8)^T$, and $\Sigma_l = \Sigma$ ($l = 1, 2$), where $\Sigma_{12} = 0.5$, $\Sigma_{11} = 1$, and $\Sigma_{22} = 2$. We performed 10 simulation replications.

This simple example demonstrated the generality of WASP in estimating the posterior distribution of functions of θ as described in Corollary 3.3. We defined two nonlinear functions of θ as

$$\rho_l = (\Sigma_l)_{12} / \{(\Sigma_l)_{11}(\Sigma_l)_{22}\}^{1/2} \quad l = 1, 2, \quad g(\mathbf{x}) = f_{\text{mix}}\{\mathbf{x}, \mathbf{x}\}^T \quad \mathbf{x} \in \mathbb{R}, \quad (13)$$

where ρ_l is the correlation of l th mixture component and $g(\mathbf{x})$ is the value of density f_{mix} in (43) when $\mathbf{y} = (\mathbf{x}, \mathbf{x})^T$. Our simulation setup implied that $\rho_1 = \rho_2$ and $g(\mathbf{x})$ was bimodal for $\mathbf{x} \in \mathbb{R}$. We completed the hierarchical model in (43) by specifying independent conjugate priors on $\boldsymbol{\pi}$ and $(\boldsymbol{\mu}_l, \Sigma_l)$ ($l = 1, 2$) as

$$\boldsymbol{\pi} \sim \text{Dirichlet}(1/2, 1/2), \quad \boldsymbol{\mu}_l | \Sigma_l \sim \mathcal{N}_2(\mathbf{0}, 100\Sigma_l), \quad \Sigma_l \sim \text{Inverse-Wishart}(2, 4I_2) \quad (l = 1, 2), \quad (14)$$

where 2 is the prior degrees of freedom and $4I_p$ is the scale matrix of the Inverse-Wishart distribution. The posterior samples of θ were obtained using Gibbs sampling (Bishop, 2006), which were used to obtain posterior samples for ρ_1 , ρ_2 , and g .

We compared WASP with the posterior distributions estimated using CMC, Gibbs sampling, SDP, and VB. We used the VB algorithm developed in Bishop (2006). Two values of $k \in \{5, 10\}$ were used for CMC, SDP, and WASP and full data were partitioned into k subsets such that the mixture proportions were preserved in every subset. The approximate posterior distributions of ρ_1 , ρ_2 , and $g(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}$, under each method were estimated using the subset posterior samples obtained after modifying the original Gibbs sampler. The sampling algorithm for WASP is described in Section 2.1 of Supplementary Material.

We compared the accuracy (11) of CMC, SDP, VB, and WASP in approximating the full data posterior distributions of ρ_1 , ρ_2 , and point-wise 90% credible bands of $g(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}$, denoted as $g_{0.05}(\mathbf{x})$ and $g_{0.95}(\mathbf{x})$. CMC, SDP, and WASP accurately approximated the full data posterior distributions of ρ_1 and ρ_2 for both k s, but VB underestimated the posterior uncertainty in ρ_1 and ρ_2 . CMC, VB, and WASP were very accurate in estimating $g_{0.05}(\mathbf{x})$ and $g_{0.95}(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}$, whereas the application of SDP failed due to a numerical error in matrix inversion (Table 1). This provides an empirical verification of Corollary 3.3, showing that the accuracy of the WASP was unaffected by the form of the parameters in the combination step. Theoretical guarantees similar to Corollary 3.3 were unavailable for CMC or SDP, but our numerical results illustrated that a similar result might also hold for these methods in mixture models.

Table 1: Accuracies of the approximate posteriors for ρ_1 , ρ_2 , and $g_{0.05}(x)$ and $g_{0.95}(x)$ for $x \in \mathbb{R}$. The accuracies are averaged over 10 simulation replications. Monte Carlo errors are in parenthesis. CMC, consensus Monte Carlo; SDP, semiparametric density product; VB, variational Bayes; WASP, Wasserstein posterior

	ρ_1		ρ_2		$g_{0.05}$		$g_{0.95}$	
VB	0.77 (0.31)		0.76 (0.29)		0.99 (0.00)		0.99 (0.00)	
	k = 5	k = 10	k = 5	k = 10	k = 5	k = 10	k = 5	k = 10
CMC	0.97 (0.01)	0.96 (0.01)	0.96 (0.01)	0.96 (0.01)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
SDP	0.97 (0.01)	0.96 (0.01)	0.95 (0.01)	0.96 (0.01)	-	-	-	-
WASP	0.97 (0.01)	0.95 (0.01)	0.97 (0.01)	0.96 (0.01)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)

4.3 Simulated data: Linear mixed effects model

Linear mixed effects models are extensively used in extending linear regression to account for longitudinal and nested dependence structures. Let n , s , and s_i be the sample size, total number of observations, and total number of observations for sample i ($i = 1, \dots, n$) so that $s = \sum_{i=1}^n s_i$. Suppose $X_i \in \mathbb{R}^{s_i \times p}$ and $Z_i \in \mathbb{R}^{s_i \times r}$ include predictors in the fixed and random effects components, respectively. Letting $\mathbf{y}_i \in \mathbb{R}^{s_i}$ be the response for sample i , the linear mixed effects model assumes that

$$\mathbf{y}_i \mid \boldsymbol{\beta}, \mathbf{u}_i, \tau^2 \sim \mathcal{N}_{s_i}(X_i \boldsymbol{\beta} + Z_i \mathbf{u}_i, \tau^2 \mathbf{I}_{n_i}), \quad \mathbf{u}_i \sim \mathcal{N}_r(\mathbf{0}, \Sigma), \quad (i = 1, \dots, n), \quad (15)$$

where $\mathbf{u}_i \in \mathbb{R}^r$ is the random effect for sample i with mean $\mathbf{0}$ and $r \times r$ covariance Σ , $\boldsymbol{\beta} \in \mathbb{R}^p$ denotes the fixed effects, and τ^2 is the error variance. The model parameters are $\theta = \{\boldsymbol{\beta}, \Sigma, \tau^2\}$.

We simulated data for a fixed n and s and varying p and r . We chose two values of $(p, r) \in \{(4, 3), (80, 6)\}$, fixed n and s to be 5000 and 100,000, and randomly assigned the s observations to n samples. The two choices of (p, r) ensured that the number of unknown parameters in $\boldsymbol{\beta}$ and Σ was 10 and 100 in the former and latter cases. The entries of X_i and Z_i were set to 1 or -1 with equal probability for every i . We fixed $\boldsymbol{\beta}$ entries as -2 and 2 alternately, $\tau^2 = 1$, and $\Sigma = LL^T$, where L was a lower triangular matrix with $L_{ii} = \sqrt{i}$ ($i = 1, \dots, r$) and the off-diagonal entries from L_{21} to $L_{r(r-1)}$ increasing linearly from $\sqrt{r-1}/100$ to $\sqrt{(r-1)}/100$. We used this setup to simulate data from (15) and performed 10 replications.

We used the HMC algorithm in Stan for sampling from the full data and subset posterior distributions. The full data posterior computations were feasible for the chosen values of n and s and posterior samples were obtained after completing the hierarchical model in (15) by using the default weakly informative priors for $\boldsymbol{\beta}$, Σ , and τ^2 in Stan. Two values of $k \in \{10, 20\}$ were used for CMC, SDP, and WASP, and the n samples were randomly partitioned into k subsets. The sampling algorithms for subset posterior distributions for the three methods were implemented in Stan and posterior samples of θ were obtained in parallel across k subsets. This was followed by a combination step to estimate the approximate posterior distributions for the three methods. The sampling algorithm for WASP is described in Section 2.2 of Supplementary Material.

We compared the accuracy (11) of CMC, SDP, VB, and WASP in approximating the marginal posterior distributions of fixed effects, variances and covariances of random effects, and the joint posterior distributions of three pairs of covariances of random effects. We used the streamlined algorithm for estimating the VB posterior for $\boldsymbol{\beta}$ and Σ (Lee and Wand, 2016). The four methods were significantly faster than the full data posterior distribution, while using only 25% of the memory resources, with VB being the fastest. CMC, SDP, VB, and WASP provided accurate approximations of the marginal posterior distributions of fixed effects and covariances of random effects. The accuracy of CMC and SDP in approximating the marginal posterior distributions of variances of random effects depended on k and r , and VB underestimated the posterior variances of random effects. In these cases, the accuracy of WASP was stable for every k and r (Tables

2 and 3). All four methods showed similar accuracies in approximating the true joint posterior distributions of three pairs of covariances of random effects. The differences in accuracies for different values of k and r were due to the differences in numerical approximation of (11) using kernel density estimation (Tables 4 and 5 and Figures 2 and 3); see Table 10 in the Supplementary Material.

The accuracy of CMC, SDP, and WASP decreased when k increased from 10 to 20 because subset posterior distributions conditioned on a smaller fraction of the data. This provided an empirical verification of Theorems 3.1 and 3.2 for the WASP. Our numerical results illustrated that a similar result might also hold for CMC and SDP. The stable performance of WASP compared to that of CMC and SDP in the approximation of the posterior distributions of variances of random effects showed that the validity of the normal approximation for subset posterior distributions was crucial in obtaining accurate approximations of full data posterior using CMC and SDP. On the other hand, WASP results were free of any such assumptions and were valid for any nonlinear function of μ and Σ ; see Corollary 3.3.

Table 2: Accuracies of the approximate posteriors for variances in (15). The accuracies are averaged over 10 simulation replications and across all diagonal elements of Σ . Monte Carlo errors are in parenthesis. VB, variational Bayes; CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	$r = 3$		$r = 6$	
VB	0.39 (0.21)		0.48 (0.24)	
	$k = 10$	$k = 20$	$k = 10$	$k = 20$
CMC	0.93 (0.04)	0.90 (0.06)	0.85 (0.06)	0.74 (0.1)
SDP	0.92 (0.05)	0.83 (0.11)	0.86 (0.08)	0.77 (0.15)
VB	0.39 (0.21)	0.39 (0.21)	0.48 (0.24)	0.48 (0.24)
WASP	0.97 (0.01)	0.97 (0.02)	0.97 (0.01)	0.96 (0.01)

Table 3: Accuracies of the approximate posteriors for covariances in (15). The accuracies are averaged over 10 simulation replications and across all off-diagonal elements of Σ . Monte Carlo errors are in parenthesis. VB, variational Bayes; CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	$r = 3$		$r = 6$	
VB	0.95 (0.01)		0.95 (0.02)	
	$k = 10$	$k = 20$	$k = 10$	$k = 20$
CMC	0.95 (0.01)	0.95 (0.02)	0.95 (0.02)	0.94 (0.03)
SDP	0.93 (0.04)	0.91 (0.06)	0.91 (0.05)	0.88 (0.08)
WASP	0.97 (0.01)	0.96 (0.02)	0.97 (0.01)	0.96 (0.01)

Table 4: Accuracies of the approximate two-dimensional joint posteriors for the covariances of random effects when $r = 3$ in (15). The accuracies are averaged over 10 simulation replications. Monte Carlo errors are in parenthesis. CMC, consensus Monte Carlo; SDP, semiparametric density product; VB, variational Bayes; WASP, Wasserstein posterior

	$(\sigma_{12}, \sigma_{13})$		$(\sigma_{12}, \sigma_{23})$		$(\sigma_{13}, \sigma_{32})$	
VB	0.94 (0.01)		0.94 (0.02)		0.94 (0.02)	
	$k = 10$	$k = 20$	$k = 10$	$k = 20$	$k = 10$	$k = 20$
CMC	0.93 (0.03)	0.93 (0.04)	0.94 (0.02)	0.92 (0.03)	0.93 (0.02)	0.92 (0.03)
SDP	0.93 (0.03)	0.93 (0.04)	0.92 (0.03)	0.91 (0.06)	0.92 (0.03)	0.90 (0.04)
WASP	0.95 (0.01)	0.95 (0.01)	0.95 (0.01)	0.94 (0.01)	0.95 (0.01)	0.95 (0.01)

Table 5: Accuracies of the approximate two-dimensional joint posteriors for the covariances of random effects when $r = 6$ in (15). The accuracies are averaged over 10 simulation replications. Monte Carlo errors are in parenthesis. CMC, consensus Monte Carlo; SDP, semiparametric density product; VB, variational Bayes; WASP, Wasserstein posterior

	$(\sigma_{12}, \sigma_{13})$		$(\sigma_{12}, \sigma_{23})$		$(\sigma_{13}, \sigma_{32})$	
VB	0.92 (0.01)		0.94 (0.01)		0.93 (0.01)	
	k = 10	k = 20	k = 10	k = 20	k = 10	k = 20
CMC	0.93 (0.03)	0.91 (0.03)	0.93 (0.02)	0.92 (0.02)	0.93 (0.02)	0.94 (0.02)
SDP	0.93 (0.03)	0.90 (0.05)	0.93 (0.02)	0.90 (0.06)	0.92 (0.03)	0.92 (0.04)
WASP	0.94 (0.02)	0.94 (0.01)	0.94 (0.01)	0.94 (0.02)	0.94 (0.01)	0.94 (0.02)

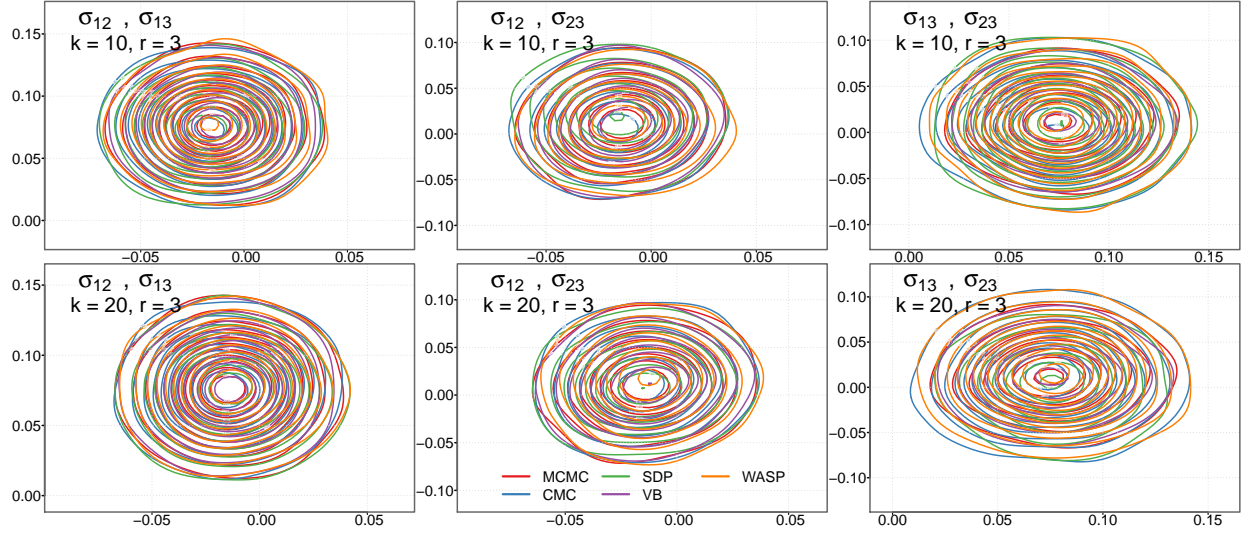


Figure 2: Kernel density estimates of the posterior densities of three covariance pairs when $r = 3$ in (15), where σ_{ab}, σ_{cd} on every panel represents the two-dimensional posterior density of $(\sigma_{ab}, \sigma_{cd})$. MCMC, Markov chain Monte Carlo; CMC, consensus Monte Carlo; SDP, semiparametric density product; VB, variational Bayes; WASP, Wasserstein posterior.

4.4 Simulated data analysis: Probabilistic parafac model

We use probabilistic parafac model as a representative example for nonparametric density estimation using WASP. Probabilistic parafac is an approach for nonparametric Bayes modeling of joint dependence in multi-variate categorical data (Dunson and Xing, 2009). Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ij}, \dots, x_{ip})$ be the data from sample i , where x_{ij} has d_j possible categorical values in $\{1, \dots, d_j\}$ ($j = 1, \dots, p$). The hierarchical model for x_{ij} ($i = 1, \dots, n; j = 1, \dots, p$) is

$$\begin{aligned}
 x_{ij} \mid \left(\psi_{h1}^{(j)} \right)_{h=1}^{\infty}, \dots, \left(\psi_{hd_j}^{(j)} \right)_{h=1}^{\infty}, z_i &\sim \text{Multinomial}(\{1, \dots, d_j\}, \psi_{z_i 1}^{(j)}, \dots, \psi_{z_i d_j}^{(j)}), \\
 z_i &\sim \sum_{h=1}^{\infty} V_h \prod_{l < h} (1 - V_l) \delta_h \equiv \sum_{h=1}^{\infty} v_h \delta_h, \quad V_h \sim \text{Beta}(1, \alpha), \\
 \psi_h^{(j)} &\sim \text{Dirichlet}(a_{j1}, \dots, a_{jd_j}), \quad \alpha \sim \text{Gamma}(a_{\alpha}, b_{\alpha}), \quad (16)
 \end{aligned}$$

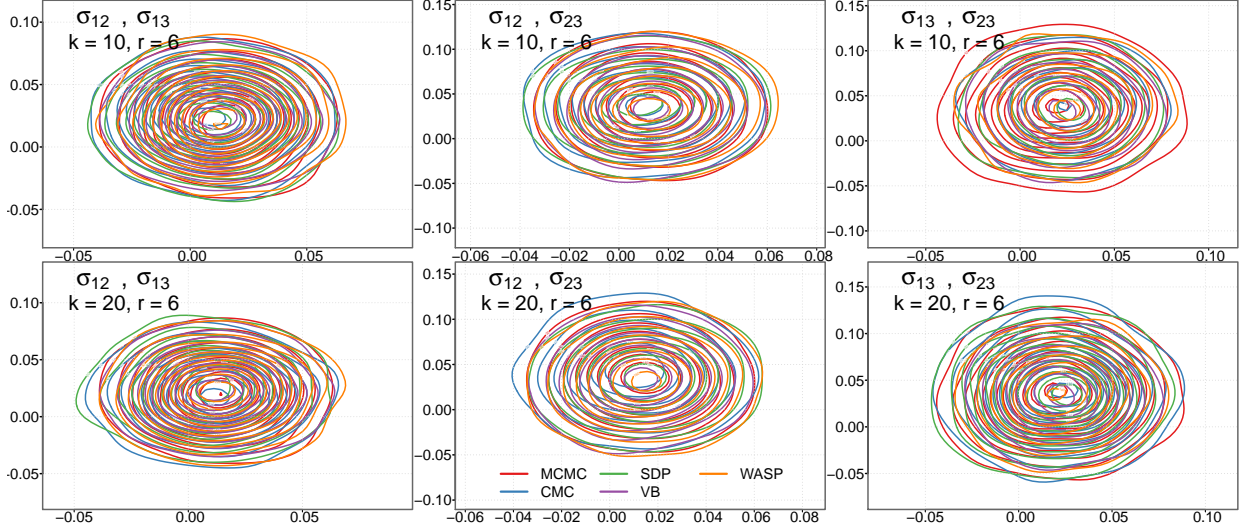


Figure 3: Kernel density estimates of the posterior densities of three covariance pairs when $r = 6$ in (15), where σ_{ab}, σ_{cd} on every panel represents the two-dimensional posterior density of $(\sigma_{ab}, \sigma_{cd})$. MCMC, Markov chain Monte Carlo; CMC, consensus Monte Carlo; SDP, semiparametric density product; VB, variational Bayes; WASP, Wasserstein posterior.

where α has prior mean a_α/b_α . The hierarchical model for probabilistic parafac implies that

$$\text{pr}(x_{i1} = c_1, \dots, x_{ij} = c_j, \dots, x_{ip} = c_p) = \pi_{c_1, \dots, c_p} = \sum_{h=1}^{\infty} v_h \prod_{j=1}^p \psi_{hc_j}^{(j)}. \quad (17)$$

The x_{ij} s are sampled independently given the latent class z_i and probability vectors $\psi_h^{(j)}$ ($h = 1, \dots, \infty$). The latent class for every sample is generated using the stick breaking representation of Dirichlet processes. The Gibbs sampling algorithm developed in Dunson and Xing (2009) is very slow even for moderate sample sizes. This example demonstrates that WASP can easily scale existing sampling algorithms to massive data, even when efficient VB alternatives are unavailable.

We followed the simulation setup in Dunson and Xing (2009), except with a much larger sample size. We fixed the sample size, number of dimensions, and number of categories in each dimension at $n = 10^5$, $p = 20$, and $d_j = 2$ ($j = 1, \dots, p$), respectively. These choices of n , p , and d_j s ensured that computations for sampling from the full data posterior were tractable. Data were simulated as a mixture of two populations such that any sample belonged to the two populations with equal probability. The two categories in every dimension excluding 2, 4, 12, and 14 were simulated from a discrete uniform in both populations. The dependence across dimensions 2, 4, 12, and 14 was induced as follows. The probabilities π_2, π_4, π_{12} , and π_{14} were set to (0.20, 0.80), (0.25, 0.75), (0.80, 0.20), and (0.75, 0.25) in the first population and to (0.80, 0.20), (0.75, 0.25), (0.20, 0.80), and (0.25, 0.75) in the second population. The simulation setup was replicated 10 times.

We used CMC, SDP, and WASP to approximate the full data posterior distributions for $\text{pr}(x_i = 1)$, where $i \in \{2, 4, 12, 14\}$. Two values of $k \in \{5, 10\}$ were used for CMC, SDP, and WASP. The full data were randomly partitioned into k subsets and subset posterior samples for WASP were obtained after modifying the Gibbs sampling algorithm in Dunson and Xing (2009) using (5). Examples for the application of CMC and SDP were unavailable for Dirichlet process mixtures, and it was unclear how to raise the prior density to the power $1/k$ when the prior distribution has an atomic form similar to that in (16); therefore, we did not raise the prior to a power of $1/k$ for sampling from the subset posterior distributions in CMC and

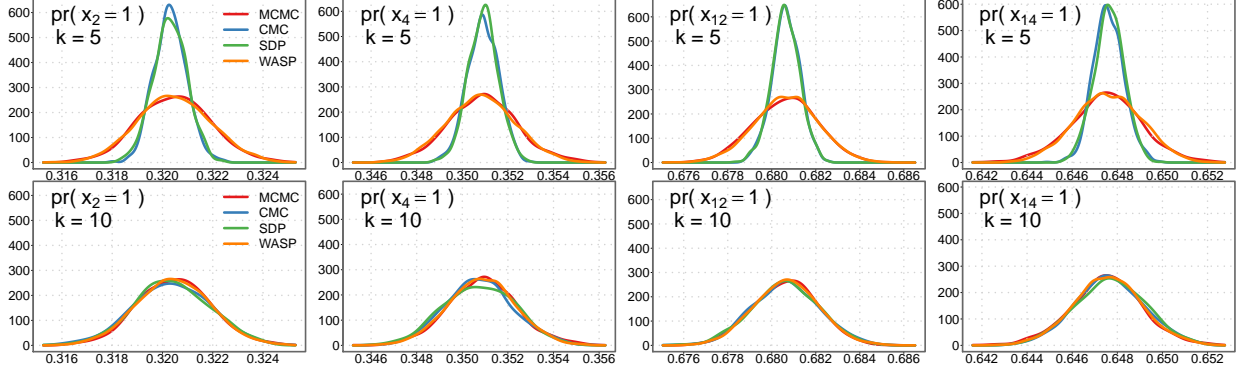


Figure 4: Kernel density estimates of the marginal posterior densities for dimensions 2, 4, 12, and 14. MCMC, Gibbs sampling algorithm of Dunson and Xing (2009); CMC, consensus Monte Carlo; SDP, semi-parametric density product; VB, variational Bayes; WASP, Wasserstein posterior

SDP. The sampling algorithm for WASP based on stochastic approximation is summarized in Section 2.3 of Supplementary Material. Subset posterior samples for $\text{pr}(x_2 = 1)$, $\text{pr}(x_4 = 1)$, $\text{pr}(x_{12} = 1)$, and $\text{pr}(x_{14} = 1)$ were combined to obtain their approximate posterior distributions using CMC, SDP, and WASP.

The accuracy (11) of CMC and SDP in approximating the full data marginal posterior distribution depended on k , with WASP outperforming CMC and SDP when $k = 5$ (Table 6). The approximate and full data posterior distributions were centered at the same value across all dimensions and replications, but the posterior densities for CMC and SDP were highly concentrated compared to the full data posterior density when $k = 5$ (Figure 4). The accuracy of WASP remained stable with varying k , providing an empirical verification of Theorems 3.1 and 3.2 in cases where our theory is not applicable. The time spent in combining subset posterior samples was negligible compared to the time spent in sampling; therefore, WASP could be used for data with much larger sample size by choosing k large enough such that sampling was efficient across all the data subsets.

Table 6: Accuracies of the approximate marginal posterior distributions for dimensions 2, 4, 12, and 14 in (16). The accuracies are averaged over 10 simulation replications. Monte Carlo errors are in parenthesis. CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	k = 5			k = 10		
	CMC	SDP	WASP	CMC	SDP	WASP
$\text{pr}(x_2 = 1)$	0.63 (0.02)	0.62 (0.02)	0.97 (0.01)	0.96 (0.02)	0.95 (0.01)	0.97 (0.01)
$\text{pr}(x_4 = 1)$	0.63 (0.02)	0.62 (0.02)	0.97 (0.01)	0.96 (0.01)	0.95 (0.02)	0.97 (0.01)
$\text{pr}(x_{12} = 1)$	0.62 (0.02)	0.62 (0.02)	0.97 (0.01)	0.95 (0.01)	0.96 (0.02)	0.97 (0.01)
$\text{pr}(x_{14} = 1)$	0.64 (0.01)	0.63 (0.01)	0.97 (0.01)	0.96 (0.02)	0.95 (0.02)	0.97 (0.01)

4.5 Real data analysis: MovieLens ratings data

We used MovieLens data to illustrate the application of WASP to large-scale ratings data. MovieLens data are one of the largest publicly available ratings data with about 10 million ratings from about 72 thousand users of the MovieLens recommender system. Each observation in the database consists of a user, movie, rating of the movie from 0.5 to 5 in increments of 0.5, and the time of rating. Every movie is also classified into at least one of the 19 genres. We fit a linear mixed effects model (15) using movie- and user-specific information as predictors and the ratings as responses.

We generated three new predictors for accurate modeling of ratings following Perry (2016). First, movie genres were grouped into *movie categories* to reduce the number of genres from 19 to four: *Action* category included Action, Adventure, Fantasy, Horror, Sci-Fi, and Thriller genres; *Children* category included Animation and Children genres; *Comedy* category included Comedy genre; and *Drama* category included Crime, Documentary, Drama, Film-Noir, Musical, Mystery, Romance, War, and Western genres. If a movie belonged to multiple genres, then movie category scores were fractions proportional to the number of genres in the respective categories. Second, *popularity* predictor was defined as $\text{logit}\{(l + 0.5)/(n + 1.0)\}$, where l and n respectively were the number of users who liked and rated the movie in 30 most recent observations for the movie and $\text{logit}(x) = \log \frac{x}{1-x}$. Third, *previous* predictor was defined to be 1 if the user liked the previous movie and 0 otherwise. We used *Action*, *Children — Action*, *Comedy — Action*, *Drama — Action*, *popularity*, and *previous* as the fixed and random effects in (15).

Following the setup in Section 4.3, we compared the performance of WASP with CMC, SDP, and VB using the full data posterior distribution as the benchmark. Sampling using the HMC algorithm in Stan was prohibitively slow for the full data posterior distribution, so we first randomly selected 5000 users and then randomly selected 20 ratings for every user. This resulted in a data set with 100,000 ratings. We randomly split the users into 10 training data sets such that ratings for any user belonged to the same training data set. To compute the approximate posteriors using CMC, SDP, and WASP, we set $k = 10$ and randomly partitioned the users into k subsets such that each subset contained all the ratings for a user. This setup was replicated for every training data.

WASP performed better than its competitors in approximating the full data posterior distributions for variances and covariances of the random effects. Similar to the simulation results in Section 4.3, CMC, SDP, VB, and WASP were significantly faster than the full data posterior distribution, with VB being the fastest. CMC, SDP, and WASP showed excellent performance in approximating the full data posterior distributions for the fixed effects. WASP outperformed its competitors in approximating the full data posterior distributions for variances, covariances, and pairs of covariances of the random effects (Tables 7, 8, and 9). VB significantly under-performed in the estimation of the posterior distribution for the fixed effects and covariance matrix of the random effects. The accuracy of marginals in CMC and SDP depended on the magnitude of covariances, with both methods showing excellent accuracy for covariances with low magnitude. The accuracies of the two-dimensional joint distributions in CMC and SDP were poor because the full data posteriors concentrated at different locations (Figure 5). Except for the poor performance of CMC, SDP, and VB in approximating the posterior distribution of variances and covariances of the random effects, our results for MovieLens data agreed with the simulation results. We concluded that WASP performed better than its competitors in the analysis of MovieLens data.

Table 7: Accuracies of the approximate posteriors for variances in (15). The accuracies are averaged over 10 replications. Monte Carlo errors are in parenthesis. CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	σ^2_{Action}	$\sigma^2_{\text{Children} - \text{Action}}$	$\sigma^2_{\text{Comedy} - \text{Action}}$	$\sigma^2_{\text{Drama} - \text{Action}}$	$\sigma^2_{\text{Popularity}}$	$\sigma^2_{\text{Previous}}$
VB	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
CMC	0.28 (0.13)	0.01 (0.01)	0.01 (0.01)	0.14 (0.09)	0.74 (0.10)	0.22 (0.10)
SDP	0.05 (0.03)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.35 (0.10)	0.03 (0.03)
WASP	0.92 (0.04)	0.93 (0.02)	0.87 (0.06)	0.85 (0.08)	0.92 (0.03)	0.93 (0.05)

5 Discussion

We have presented WASP as an approach for computationally efficient approximation of the posterior distributions of parameters and their functions when the sample size is large. WASP allows extensions of existing

Table 8: Accuracies of the approximate posteriors for covariances in (15). The accuracies are averaged over 10 replications. Monte Carlo errors are in parenthesis. The subscripts 1, . . . , 6 are used for predictors *Action*, *Children — Action*, *Comedy — Action*, *Drama — Action*, *popularity*, and *previous*. CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	σ_{12}	σ_{13}	σ_{14}	σ_{15}	σ_{16}	σ_{23}	σ_{24}	σ_{25}
VB	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
CMC	0.06 (0.03)	0.16 (0.04)	0.18 (0.04)	0.83 (0.07)	0.33 (0.13)	0.01 (0.01)	0.07 (0.02)	0.80 (0.04)
SDP	0.01 (0.01)	0.08 (0.03)	0.07 (0.02)	0.75 (0.06)	0.14 (0.09)	0.00 (0.00)	0.02 (0.01)	0.73 (0.08)
WASP	0.95 (0.02)	0.91 (0.04)	0.91 (0.05)	0.94 (0.03)	0.90 (0.07)	0.89 (0.07)	0.85 (0.08)	0.93 (0.03)
	σ_{26}	σ_{34}	σ_{35}	σ_{36}	σ_{45}	σ_{46}	σ_{56}	
VB	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	
CMC	0.66 (0.09)	0.65 (0.07)	0.76 (0.08)	0.71 (0.05)	0.82 (0.04)	0.61 (0.11)	0.55 (0.09)	
SDP	0.59 (0.11)	0.62 (0.06)	0.64 (0.09)	0.66 (0.08)	0.66 (0.09)	0.56 (0.14)	0.55 (0.13)	
WASP	0.91 (0.05)	0.94 (0.05)	0.93 (0.03)	0.91 (0.04)	0.93 (0.04)	0.93 (0.04)	0.94 (0.04)	

Table 9: Accuracies of the approximate two-dimensional joint posteriors for the covariances of random effects. The accuracies are averaged over 10 replications. Monte Carlo errors are in parenthesis. The subscripts 1, . . . , 6 are used for predictors *Action*, *Children — Action*, *Comedy — Action*, *Drama — Action*, *popularity*, and *previous*. CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	$(\sigma_{12}, \sigma_{13})$	$(\sigma_{12}, \sigma_{14})$	$(\sigma_{12}, \sigma_{15})$	$(\sigma_{12}, \sigma_{16})$
VB	0.18 (0.04)	0.22 (0.07)	0.31 (0.03)	0.31 (0.02)
CMC	0.05 (0.02)	0.04 (0.02)	0.06 (0.03)	0.05 (0.02)
SDP	0.05 (0.02)	0.04 (0.02)	0.06 (0.03)	0.05 (0.02)
WASP	0.88 (0.03)	0.88 (0.03)	0.88 (0.02)	0.86 (0.06)

samplers to massive data with minimal modifications and is easily implemented using probabilistic programming languages, such as Stan. Theoretically, we showed that the rate of convergence of WASP to the delta measure centered at the true parameter value in W_2 distance matches the optimal parametric rate up to a logarithmic factor if the number of subsets increases slowly with the size of the full dataset. Empirically, we demonstrated that results from WASP and MCMC agree closely in several widely different examples, while WASP enables massive speed-ups in computational time.

We plan to explore several extensions of WASP in the future. First, it is unclear how to extend stochastic approximation to cases where the likelihood is unavailable in a product form. This extension is crucial for proper uncertainty quantification outside of settings in which the observations are conditionally independent given latent variables. Second, it is unclear how to optimally choose k in practice; larger k improves computational time when abundant processors are available but choosing k too large may lead to increasing statistical errors (refer to Theorem 3.1). Our numerical experiments show that the accuracy of WASP is robust to the choice of k if all the subset sizes are moderately large relative to the number of parameters. In addition, it is of interest to study more deeply the impact of the partitioning schemes and attempt to develop approaches that deal with not only large sample sizes but also high-dimensional data. A possibility in this regard is to combine WASP with approximate MCMC (Johndrow et al., 2015).

Acknowledgment

We thank Volkan Cevher and Quoc Tran-Dinh for proposing and implementing the algorithm for calculating Wasserstein barycenter described in Srivastava et al. (2015). All experiments were based on a modified version of their Matlab and Gurobi code for estimating Wasserstein barycenter.

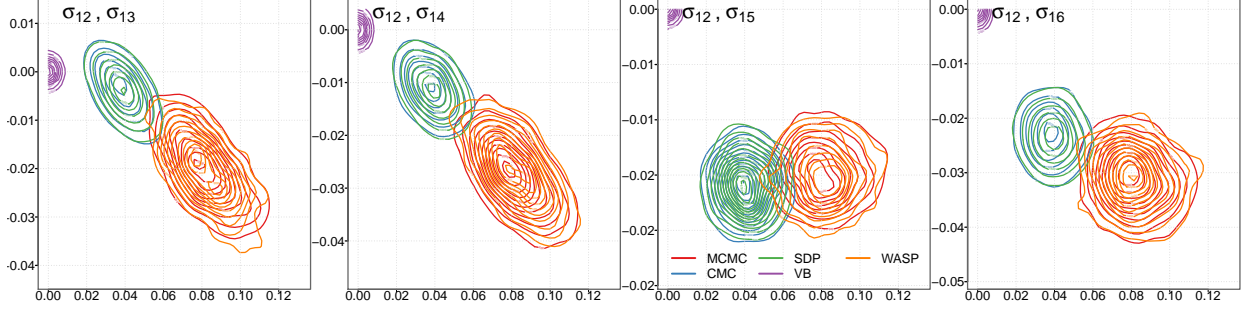


Figure 5: Kernel density estimates of the posterior densities of four covariance pairs, where σ_{ab}, σ_{cd} on every panel represents the two-dimensional posterior density of $(\sigma_{ab}, \sigma_{cd})$. MCMC, Markov chain Monte Carlo; CMC, consensus Monte Carlo; SDP, semiparametric density product; VB, variational Bayes; WASP, Wasserstein posterior.

Appendix

A Assumptions

Suppose $h(p_1, p_2) = [\int (\sqrt{p_1} - \sqrt{p_2})^2 d\nu]^{1/2}$ is the Hellinger distance between two generic densities p_1, p_2 . We define the following pseudo Hellinger distance on each subset of data $Y_{[j]}$, with sample size m and $j = 1, \dots, k$.

Definition A.1 (Pseudo Hellinger distance) *The pseudo Hellinger (h_{mj}) distance between $P_{\theta_1}^{[j]}$ and $P_{\theta_2}^{[j]}$ in $\{\otimes_{i=1}^m P_{\theta, j, i} : \theta \in \Theta\}$ is defined as*

$$h_{mj}^2(\theta_1, \theta_2) = \frac{1}{m} \sum_{i=1}^m h^2(p_{ji}(\cdot | \theta_1), p_{ji}(\cdot | \theta_2)). \quad (18)$$

This definition is based on Equation (3.1) in Ghosal and van Der Vaart (2007) and it is straightforward to check that for every $j = 1, \dots, k$, $(\{\otimes_{i=1}^m P_{\theta, j, i} : \theta \in \Theta\}, h_{mj})$ is a metric space.

For a m -dimensional random vector $Z = (Z_1, \dots, Z_m)^T$, denote its L_q norm as

$$|Z|_q = \left[\frac{1}{m} \sum_{i=1}^m \mathbb{E}(|Z_i|^q) \right]^{1/q},$$

and we also use $\|Z\|$ to represent $|Z|_2$.

Next we define the *generalized bracketing entropy* for *inid* data. If the data are indeed *iid* then the generalized bracketing entropy coincides with the usual bracketing entropy.

Definition A.2 (Generalized bracketing entropy) *Let Ξ be a fixed subset of Θ . For a fixed $j \in \{1, \dots, k\}$, let*

$$\mathcal{P}_j(\Xi) = \{p_j(\mathbf{y}|\theta) = (p_{j1}(y_1|\theta), \dots, p_{jm}(y_m|\theta))^T : \mathbf{y} = (y_1, \dots, y_m)^T \in \otimes_{i=1}^m \mathcal{Y}_{ji}, \theta \in \Xi\}$$

be the class of m -dimensional functions indexed by θ . For a given $\delta > 0$, let

$$\mathcal{B}(\delta, \mathcal{P}_j(\Xi)) = \left\{ [\mathbf{l}_s, \mathbf{u}_s] : \mathbf{l}_s(\mathbf{y}) = (l_{s1}(y_1), \dots, l_{sm}(y_m))^T, \mathbf{u}_s(\mathbf{y}) = (u_{s1}(y_1), \dots, u_{sm}(y_m))^T, \right.$$

$$\mathbf{y} = (y_1, \dots, y_m)^T \in \otimes_{i=1}^m \mathcal{Y}_{ji}, s = 1, \dots, N \}$$

be the generalized bracketing set of $\mathcal{P}_j(\Xi)$ with cardinality N , such that for any $\mathbf{p}_j(\mathbf{y}|\theta) \in \mathcal{P}_j(\Xi)$, there exists a pair of functions $[\mathbf{l}_s, \mathbf{u}_s] \in \mathcal{B}(\delta, \mathcal{P}_j(\Xi))$, such that

$$\begin{aligned} l_{si}(y_i) &\leq p_{ji}(y_i) \leq u_{si}(y_i), \text{ for all } \mathbf{y} \in \otimes_{i=1}^m \mathcal{Y}_{ji}, \text{ and all } i = 1, \dots, m \\ \text{and } \|\sqrt{\mathbf{u}_s} - \sqrt{\mathbf{l}_s}\| &\leq \delta. \end{aligned}$$

The h_{mj} -bracketing number of $\mathcal{P}_j(\Xi)$, $N_{[]}(\delta, \mathcal{P}_j(\Xi), h_{mj})$, is defined as the smallest cardinality of the generalized bracketing set $\mathcal{B}(\delta, \mathcal{P}_j(\Xi))$. The h_{mj} -bracketing entropy of $\mathcal{P}_j(\Xi)$ is defined as $H_{[]}(\delta, \mathcal{P}_j(\Xi), h_{mj}) = \log(1 + N_{[]}(\delta, \mathcal{P}_j(\Xi), h_{mj}))$.

B Proofs of Theorems

Let $\epsilon_m = \left(\frac{m}{\log^2 m}\right)^{-1/(2\alpha)}$. For ease of notation, in all the following proofs, we will sometimes write $p(y_{ji} | \theta) \equiv p_{ji}(y_{ji} | \theta)$.

B.1 Proof of Theorem 3.1

Due to the compactness of Θ in (A1), we assume that $\rho(\theta, \theta_0) \leq M_0$ for a large finite constant M_0 . We start with a decomposition of the W_2 distance from the j th subset posterior $\Pi_m^\gamma(\cdot | Y_{[j]})$ to the delta measure at the true parameter θ_0 :

$$\begin{aligned} E_{P_{\theta_0}} W_2^2(\Pi_m^\gamma(\cdot | Y_{[j]}), \delta_{\theta_0}(\cdot)) &= E_{P_{\theta_0}} \int_{\Theta} \rho^2(\theta, \theta_0) \Pi_m^\gamma(d\theta | Y_{[j]}) \\ &\leq E_{P_{\theta_0}} \int_{\{\theta: \rho(\theta, \theta_0) \leq c_1 \epsilon_m\}} \rho^2(\theta, \theta_0) \Pi_m^\gamma(d\theta | Y_{[j]}) + E_{P_{\theta_0}} \int_{\{\theta: \rho(\theta, \theta_0) > c_1 \epsilon_m\}} \rho^2(\theta, \theta_0) \Pi_m^\gamma(d\theta | Y_{[j]}) \\ &\leq (c_1 \epsilon_m)^2 + M_0^2 E_{P_{\theta_0}} \Pi_m^\gamma(\rho(\theta, \theta_0) > c_1 \epsilon_m | Y_{[j]}). \end{aligned} \quad (19)$$

We will choose the constant c_1 as $c_1 = \left(\frac{2r_1 g_2}{q_1 C_L}\right)^{1/(2\alpha)}$, where g_1, C_L, q_1, r_1 are the constants in (A1), (A2), and Lemma 1.5 and Lemma 1.6 in Supplementary Material.

The following proofs are similar to the proofs of Theorem 1, 4, and 10 in Ghosal and van Der Vaart (2007). The main difference is that our likelihood has been raised to the power γ . Using condition (A2), we can further replace the ρ metric by the pseudo Hellinger distance:

$$\begin{aligned} &\Pi_m^\gamma(\theta \in \Theta : \rho(\theta, \theta_0) > c_1 \epsilon_m | Y_{[j]}) \\ &\leq \Pi_m^\gamma(\theta \in \Theta : h_{mj}(P_{\theta, j}, P_{\theta_0, j}) > \sqrt{C_L} (c_1 \epsilon_m)^\alpha | Y_{[j]}) \\ &= \int_{\{\theta \in \Theta : h_{mj}(\theta, \theta_0) > \sqrt{\frac{2r_1 g_2}{q_1}} \epsilon_m^\alpha\}} \frac{\prod_{i=1}^m \left[\frac{p(Y_{ji}|\theta)}{p(Y_{ji}|\theta_0)}\right]^\gamma \Pi(d\theta)}{\int_{\Theta} \prod_{i=1}^m \left[\frac{p(Y_{ji}|\theta)}{p(Y_{ji}|\theta_0)}\right]^\gamma \Pi(d\theta)}. \end{aligned} \quad (20)$$

For the denominator in (20), by Condition (A4) and Lemma 1.6, for m sufficiently large, with probability at least $1 - \exp(-r_2 m \epsilon_m^{2\alpha})$

$$\int_{\Theta} \prod_{i=1}^m \frac{p(Y_{ji}|\theta)^\gamma}{p(Y_{ji}|\theta_0)^\gamma} \Pi(d\theta) > \exp(-r_1 n \epsilon_m^{2\alpha}). \quad (21)$$

For the numerator in (20), by Condition (A3) and Lemma 1.5, we set $\delta = \sqrt{2r_1 g_2 / q_1} \epsilon_m^\alpha$ and obtain that with probability at least $1 - 4 \exp\left(-\frac{2r_1 q_2 q_2}{q_1} m \epsilon_m^{2\alpha}\right) \geq 1 - 4 \exp\left(-\frac{2r_1 q_2}{q_1} n \epsilon_m^{2\alpha}\right)$,

$$\sup_{\{\theta \in \Theta : h_{mj}(\theta, \theta_0) \geq \sqrt{2r_1 g_2 / q_1} \epsilon_m^\alpha\}} \prod_{i=1}^m \left[\frac{p(Y_{ji}|\theta)}{p(Y_{ji}|\theta_0)} \right]^\gamma \leq \exp(-2r_1 g_2 m \epsilon_m^{2\alpha}) \leq \exp(-2r_1 n \epsilon_m^{2\alpha}) \quad (22)$$

Therefore, based on (20), (21), and (22), we obtain that with probability at least $1 - 4 \exp\left(-\frac{2r_1 q_2}{q_1} n \epsilon_m^{2\alpha}\right) - \exp(-r_2 m \epsilon_m^{2\alpha})$,

$$\Pi_m^\gamma \left(\theta \in \Theta : \rho(\theta, \theta_0) > c_1 \epsilon_m \mid Y_{[j]} \right) \leq \exp(-2r_1 n \epsilon_m^{2\alpha} + r_1 n \epsilon_m^{2\alpha}) \leq \exp(-r_1 n \epsilon_m^{2\alpha}).$$

Let A_{ϵ_m} be the event $\{\theta \in \Theta : \Pi \left(\theta \in \Theta : \rho(\theta, \theta_0) > c_1 \epsilon_m \mid Y_{[j]} \right) \leq \exp(-r_1 n \epsilon_m^{2\alpha})\}$. Then we can bound the second term in (19) as

$$\begin{aligned} & E_{P_{\theta_0}} \Pi_m^\gamma \left(\rho(\theta, \theta_0) > c_1 \epsilon_m \mid Y_{[j]} \right) \\ & \leq E_{P_{\theta_0}} \left[I(A_{\epsilon_m}) \Pi_m^\gamma \left(\rho(\theta, \theta_0) > c_1 \epsilon_m \mid Y_{[j]} \right) \right] + E_{P_{\theta_0}} \left[I(A_{\epsilon_m}^c) \Pi_m^\gamma \left(\rho(\theta, \theta_0) > c_1 \epsilon_m \mid Y_{[j]} \right) \right] \\ & \leq \exp(-r_1 n \epsilon_m^{2\alpha}) + P_{\theta_0}^{(n)}(A_{\epsilon_m}^c) \cdot 1 \\ & \leq \exp(-r_1 n \epsilon_m^{2\alpha}) + 4 \exp\left(-\frac{2r_1 q_2}{q_1} n \epsilon_m^{2\alpha}\right) + \exp(-r_2 m \epsilon_m^{2\alpha}) \\ & \leq 6 \exp(-c_2 m \epsilon_m^{2\alpha}), \end{aligned}$$

for $c_2 = \min(r_1, r_2, 2r_1 q_2 / q_1)$, as clearly the second term is dominating the other two given $m \lesssim n$.

Therefore, for (19), since $\epsilon_m = (m / \log^2 m)^{-1/(2\alpha)}$, as $m \rightarrow \infty$, an explicit bound will be

$$\begin{aligned} & E_{P_{\theta_0}} W_2^2 \left(\Pi_m^\gamma(\cdot \mid Y_{[j]}), \delta_{\theta_0}(\cdot) \right) \leq c_1^2 \frac{\log^{2/\alpha} m}{m^{1/\alpha}} + 6M_0^2 \exp(-c_2 \log^2 m) \\ & \leq c_1^2 \frac{\log^{2/\alpha} m}{m^{1/\alpha}} + \frac{1}{m^{1+\frac{1}{\alpha}}} \leq C_1 \frac{\log^{2/\alpha} m}{m^{1/\alpha}} \end{aligned}$$

as m becomes sufficiently large, where the constant C_1 depends on α, c_1, c_2 , which further depends on $g_1, g_2, q_1, q_2, r_1, r_2, C_L$. Since q_1, q_2 in Lemma 1.5 and r_1, r_2 in Lemma 1.6 depend on $g_1, g_2, D_1, D_2, \kappa, c_\pi$, it follows that C_1 depends on $g_1, g_2, C_L, D_1, D_2, \kappa, c_\pi$. \square

B.2 Proof of Theorem 3.2

Based on Theorem 3.1 and Lemma 1.7 in Supplementary Material, we have that for any $C > 0$,

$$\begin{aligned} & P_{\theta_0}^{(n)} \left(W_2 \left(\bar{\Pi}_n^\gamma(\cdot \mid Y^{(n)}), \delta_{\theta_0}(\cdot) \right) > \sqrt{\frac{C \log^{2/\alpha} m}{m^{1/\alpha}}} \right) \\ & \leq P_{\theta_0}^{(n)} \left(\frac{1}{k} \sum_{j=1}^k W_2 \left(\Pi_m^\gamma(\cdot \mid Y_{[j]}), \delta_{\theta_0}(\cdot) \right) > \sqrt{\frac{C \log^{2/\alpha} m}{m^{1/\alpha}}} \right) \\ & \stackrel{(i)}{\leq} \frac{1}{\frac{C \log^{2/\alpha} m}{m^{1/\alpha}}} E_{P_{\theta_0}} \left[\frac{1}{k} \sum_{j=1}^k W_2 \left(\Pi_m^\gamma(\cdot \mid Y_{[j]}), \delta_{\theta_0}(\cdot) \right) \right]^2 \end{aligned}$$

$$\begin{aligned}
&\stackrel{(ii)}{\leq} \frac{m^{1/\alpha}}{Ck \log^{2/\alpha} m} \sum_{j=1}^k \mathbb{E}_{P_{\theta_0}} W_2^2(\Pi_m^\gamma(\cdot | Y_{[j]}), \delta_{\theta_0}(\cdot)) \\
&\leq \frac{m^{1/\alpha}}{Ck \log^{2/\alpha} m} \cdot k C_1 \frac{\log^{2/\alpha} m}{m^{1/\alpha}} = \frac{C_1}{C},
\end{aligned}$$

where we have used Markov's inequality in (i) and the relation between l_1 and l_2 norms in (ii). This implies that $W_2(\bar{\Pi}_n^\gamma(\cdot | Y^{(n)}), \delta_{\theta_0}(\cdot)) = O_p\left(\sqrt{\frac{\log^{2/\alpha} m}{m^{1/\alpha}}}\right)$. \square

C Univariate density estimation

Let X_1, \dots, X_n be n copies of a scalar random variable X that follows probability distribution P_0 with density p_0 . The full data are randomly split into k subsets and X_{j1}, \dots, X_{jm} represent the data on subset j ($j = 1, \dots, k$). The hierarchical model for density estimation using the stick-breaking representation of Dirichlet process mixtures is

$$\begin{aligned}
X_{ji} | z_{ji}, \{\mu_h\}_{h=1}^\infty, \{\sigma_h^2\}_{h=1}^\infty &\sim \mathcal{N}(\mu_{z_{ji}}, \sigma_{z_{ji}}^2), \quad z_{ji} \sim \sum_{h=1}^\infty v_h \delta_h, \quad v_h = V_h \prod_{l < h} (1 - V_l), \quad V_h | \alpha \sim \text{Beta}(1, \alpha), \\
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha), \quad \mu_h | \sigma_h^2 \sim \mathcal{N}(0, \sigma_h^2), \quad \sigma_h^2 \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma),
\end{aligned} \tag{23}$$

where $a_\sigma > 2$ and Beta, Gamma, and Inverse-Gamma random variables have means $\frac{1}{1+\alpha}$, $\frac{a_\alpha}{b_\alpha}$, and $\frac{b_\sigma}{a_\sigma-1}$ and variances $\frac{\alpha}{(1+\alpha)^2(2+\alpha)}$, $\frac{a_\alpha}{b_\alpha^2}$, and $\frac{b_\sigma^2}{(a_\sigma-1)^2(a_\sigma-2)}$. If l^* is the maximum number of atoms in the stick-breaking representation, then the prior density π is in the form a discrete mixture. We cannot use existing sampling algorithms directly if π is raised to a power of $1/k$, so it is unclear how to sample from the subset posterior density of competing approaches in Section 2.2.

We show that it is still possible to sample from the subset posterior density in (5) using data augmentation. Let L_j be the likelihood given X_{j1}, \dots, X_{jm} and latent variables z_{j1}, \dots, z_{jm} in (23), then

$$L_j(\{\mu_h\}_{h=1}^{l^*}, \{\sigma_h^2\}_{h=1}^{l^*}, \{v_h\}_{h=1}^{l^*}) = \prod_{h=1}^{l^*} (2\pi\sigma_h^2)^{-\frac{\#h_j}{2}} e^{-\frac{1}{2\sigma_h^2} \sum_{i=1}^m 1(z_{ji}=h) (x_{ji}-\mu_h)^2} v_h^{\#h_j}, \tag{24}$$

where $1(z_{ji} = h)$ is 1 if $z_{ji} = h$ and 0 otherwise and $\#h_j = \sum_{i=1}^m 1(z_{ji} = h)$. For stochastic approximation, we raise L_j in (24) to the power γ and obtain

$$L_j^\gamma(\{\mu_h\}_{h=1}^{l^*}, \{\sigma_h^2\}_{h=1}^{l^*}, \{v_h\}_{h=1}^{l^*}) = \prod_{h=1}^{l^*} (2\pi\sigma_h^2)^{-\frac{\gamma\#h_j}{2}} e^{-\frac{\gamma}{2\sigma_h^2} \sum_{i=1}^m 1(z_{ji}=h) (x_{ji}-\mu_h)^2} v_h^{\gamma\#h_j}. \tag{25}$$

Standard arguments imply that the analytic form of full conditional densities of parameters are

$$\begin{aligned}
\mu_h | \text{rest} &\propto e^{-\frac{\gamma\#h_j+1}{2\sigma_h^2}} \left(\mu_h^2 - 2\mu_h \gamma \frac{\sum_{i=1}^m 1(z_{ji}=h) x_{ji}}{\gamma\#h_j+1} \right), \\
\sigma_h^2 | \text{rest} &\propto \sigma_h^{2-\frac{\gamma\#h_j}{2}} e^{-\frac{\gamma}{2\sigma_h^2} \sum_{i=1}^m 1(z_{ji}=h) (x_{ji}-\mu_h)^2} \sigma_h^{-\frac{1}{2}} e^{-\frac{\mu_h^2}{2\sigma_h^2}} \sigma_h^{-a_\sigma-1} e^{-\frac{b_\sigma}{\sigma_h^2}}, \\
V_h | \text{rest} &\propto V_h^{\sum_{i=1}^m 1(z_{ji}=h)} (1 - V_h)^\gamma \sum_{i=1}^m 1(z_{ji} > h) (1 - V_h)^{\alpha-1}, \\
\alpha | \text{rest} &\propto \alpha^{a_\alpha-1} e^{-b_\alpha \alpha} \alpha^{l^*} \prod_{h=1}^{l^*} (1 - V_d)^{\alpha-1}
\end{aligned} \tag{26}$$

for $h = 1, \dots, l^*$. Let

$$m_{jh} = \frac{\gamma \sum_{i=1}^m 1(z_{ji} = h) x_{ji}}{\gamma \#h_j + 1}, \quad v_{jh} = \frac{\sigma_h^2}{\gamma \#h_j + 1}, \quad (27)$$

$$a_{jh} = \frac{\gamma \#h_j + 1}{2} + a_\sigma, \quad b_{jh} = \frac{\gamma}{2} \sum_{i=1}^m 1(z_{ji} = h) (x_{ji} - \mu_h)^2 + \frac{\mu_h^2}{2} + b_\sigma \quad (28)$$

for $h = 1, \dots, l^*$, then all full conditional densities are tractable in terms of standard distributions:

$$\begin{aligned} \mu_{jh} | \text{rest} &\sim \mathcal{N}(m_{jh}, v_{jh}), \quad \sigma_{jh}^2 | \text{rest} \sim \text{Inverse-Gamma}(a_{jh}, b_{jh}), \\ V_{jh} | \text{rest} &\sim \text{Beta}(1 + \gamma \sum_{i=1}^m 1(z_{ji} = h), \alpha + \gamma \sum_{i=1}^m 1(z_{ji} > h)), \\ \alpha_{jh} | \text{rest} &\sim \text{Gamma}(a_\alpha + l^*, b_\alpha - \sum_{h=1}^{l^*} \log(1 - V_{jh})). \end{aligned} \quad (29)$$

Finally, posterior distribution of the latent variables is

$$z_{ji} | \text{rest} \sim \sum_{h=1}^{l^*} p_{jh} \delta_h, \quad p_{jh} = \frac{v_{jh} \mathcal{N}(\mu_{jh}, \sigma_{jh}^2)}{\sum_{\tilde{h}=1}^{l^*} v_{j\tilde{h}} \mathcal{N}(\mu_{j\tilde{h}}, \sigma_{j\tilde{h}}^2)}, \quad (i = 1, \dots, m), \quad (30)$$

where $v_{jh} = V_{jh} \prod_{l < h} (1 - V_{jl})$ and $\mathcal{N}(m, v)$ is the Gaussian density with mean m and variance v .

D Linear program

$$\begin{aligned} &\underset{\mathbf{a}, T_1, \dots, T_k}{\text{minimize}} && \sum_{j=1}^k \text{trace}(T_j^T D_j) \\ &\text{subject to} && 0 \leq a_i \leq 1, \quad i = 1, \dots, g, \\ & && 0 \leq (T_j)_{uv} \leq 1, \quad u = 1, \dots, g, \quad v = 1, \dots, s_j, \quad j = 1, \dots, k, \\ & && \mathbf{1}^T \mathbf{a} = 1, \\ & && T_j \mathbf{1}_{s_j} = \mathbf{a}, \quad j = 1, \dots, k, \\ & && T_j^T \mathbf{1}_s = \frac{\mathbf{1}_{s_j}}{s_j}, \quad j = 1, \dots, k. \end{aligned} \quad (31)$$

This linear program can be solved using a variety of linear programming solvers in `Matlab` or `R`. More specialized algorithms are developed in Cuturi and Doucet (2014), Carlier et al. (2015), and Srivastava et al. (2015). We provide a simple implementation using `Matlab` in Section 3 of Supplementary Material.

Supplementary Materials for Scalable Bayes via Barycenter in Wasserstein Space

1 Technical Lemmas

To show Theorem 3.1, we first introduce in Lemma 1.5 a generalized version of the concentration inequality in Theorem 1 of Wong and Shen (1995). The proof parallels the original proof in Wong and Shen (1995), with several adaptations for the *inid* setup. Let $Z_{ji}(\theta) = \log[p(Y_{ji}|\theta)/p(Y_{ji}|\theta_0)]$, and $\tilde{Z}_{ji}(\theta) = \max(Z_{ji}(\theta), -\tau)$ be the lower truncated version of $Z_{ji}(\theta)$ for some constant $\tau > 0$ to be chosen later. Let $Z_j(\theta) = (Z_{j1}(\theta), \dots, Z_{jm}(\theta))^T$ and $\tilde{Z}_j(\theta) = (\tilde{Z}_{j1}(\theta), \dots, \tilde{Z}_{jm}(\theta))^T$.

Lemma 1.1 *Let $c_{1\tau} = 2e^{-\tau/2}/(1 - e^{-\tau/2})^2$. Then for any $\theta \in \Theta$,*

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E} \tilde{Z}_{ji}(\theta) \leq -(1 - c_{1\tau}) h_{mj}^2(\theta, \theta_0). \quad (32)$$

Proof of Lemma 1.1:

The proof is a simple adaptation of Lemma 2 and Lemma 4 in Wong and Shen (1995). First note that for every observation Y_{ji} (which takes value y_{ji}), we can define the event $A_{ji} = \{y_{ji} : p(y_{ji}|\theta)/p(y_{ji}|\theta_0) < e^{-\tau}\}$. Their Lemma 2 implies that $P(A_{ji}) \leq (1 - e^{-\tau/2})^{-2} h^2(p_{ji}(\cdot|\theta), p_{ji}(\cdot|\theta_0))$, which further implies by simple averaging over $i = 1, \dots, m$ that

$$\frac{1}{m} \sum_{i=1}^m P(A_{ji}) \leq (1 - e^{-\tau/2})^{-2} h_{mj}^2(\theta, \theta_0). \quad (33)$$

Following the same derivation of their Lemma 4, we have for every individual \tilde{Z}_{ji} ,

$$\mathbb{E} \tilde{Z}_{ji}(\theta) \leq -h^2(p_{ji}(\cdot|\theta), p_{ji}(\cdot|\theta_0)) + 2e^{-\tau/2} P(A_{ji}).$$

Then a simple averaging over $i = 1, \dots, m$ together with (33) gives (32). \square

Lemma 1.2 *Let $c_{2\tau} = (e^{\tau/2} - 1 - \tau/2)/(1 - e^{-\tau/2})^2$. For any $t > 0$, integer $\ell \geq 2$ and any $\theta \in \Theta$ that satisfies $h_{mj}(\theta, \theta_0) \leq r$,*

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}_{P_{\theta_0}} \left| \frac{\tilde{Z}_{ji}(\theta)}{2\sqrt{2c_{2\tau}r}} \right|^\ell \leq \frac{\ell!}{2} \left(\frac{1}{\sqrt{2c_{2\tau}r}} \right)^{\ell-2}.$$

Proof of Lemma 1.2:

Lemma 5 in Wong and Shen (1995) is stated for every single observation Y_{ji} , so it still holds for individual $\tilde{Z}_{ji}(\theta)$. For all $i = 1, \dots, m$,

$$\mathbb{E}_{P_{\theta_0}} \left[\exp \left(\left| \tilde{Z}_{ji}(\theta)/2 \right| \right) - 1 - \left| \tilde{Z}_{ji}(\theta)/2 \right| \right] \leq c_{2\tau} h^2(p_{ji}(\cdot|\theta), p_{ji}(\cdot|\theta_0)),$$

where $c_{2\tau}$ is as defined in the statement of the lemma. Averaging over $i = 1, \dots, m$ yields

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}_{P_{\theta_0}} \left[\exp \left(\left| \tilde{Z}_{ji}(\theta)/2 \right| \right) - 1 - \left| \tilde{Z}_{ji}(\theta)/2 \right| \right] \leq c_{2\tau} h_{mj}^2(\theta, \theta_0) \leq c_{2\tau} r^2,$$

where the second inequality is from the condition $h_{mj}(\theta, \theta_0) \leq r$. Using $e^x - 1 - x \geq x^\ell/\ell!$ for all $x > 0$, we have

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}_{P_{\theta_0}} \left| \tilde{Z}_{ji}(\theta) \right|^\ell \leq 2^\ell \ell! c_{2\tau} r.$$

Rearranging the terms and the conclusion follows. \square

Lemma 1.3 *Let $j \in \{1, \dots, k\}$ be fixed. Suppose Ξ is a subset of Θ . Let $\tilde{Z}_j(\Xi) = \{\tilde{Z}_j(\theta), \theta \in \Xi\}$. For any $u > 0$,*

$$H_\square(u, \tilde{Z}_j(\Xi), \|\cdot\|) \leq H_\square(u/(2e^{\tau/2}), \mathcal{P}_j(\Xi), h_{mj}).$$

Proof of Lemma 1.3:

The proof follows the argument in the proof of Lemma 3 in Wong and Shen (1995). We can derive that for each $i = 1, \dots, m$, for any $\theta_1, \theta_2 \in \Xi$,

$$\mathbb{E}_{P_{\theta_0}} \left[\tilde{Z}_{ji}(\theta_1) - \tilde{Z}_{ji}(\theta_2) \right]^2 \leq 4e^\tau h_{mj}^2(\theta_1, \theta_2).$$

Then averaging over $i = 1, \dots, m$ gives the relation between two norms

$$\left\| \tilde{Z}_j(\theta_1) - \tilde{Z}_j(\theta_2) \right\| \leq 2e^{\tau/2} h_{mj}(\theta_1, \theta_2),$$

which further implies the relation between the bracketing entropies. \square

Lemma 1.4 (van der Geer and Lederer (2013) Theorem 8) *Let $j \in \{1, \dots, k\}$ be fixed. Suppose a class of functions $\mathcal{F}_j = \{f(\mathbf{y}) = (f_1(\mathbf{y}_1), \dots, f_m(\mathbf{y}_m))^T, \mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_m) \in \otimes_{i=1}^m \mathcal{Y}_{ji}\}$ satisfies*

(i) $\sup_{f \in \mathcal{F}_j} \|f\| \leq 1$;

(ii) *For any integer $\ell \geq 2$, $\sup_{f \in \mathcal{F}_j} |f|_\ell^\ell \leq \ell! M^{\ell-2}/2$, for some constant $M > 0$;*

Then for any $t > 0$,

$$P_{\theta_0} \left(\sup_{f \in \mathcal{F}_j} \frac{1}{\sqrt{m}} \sum_{i=1}^m \left[f_i(Y_{ji}) - \mathbb{E}_{P_{\theta_0}} f_i(Y_{ji}) \right] \geq \min_{S \in \mathcal{N}} R_S + \frac{36M(1+t)}{\sqrt{m}} + 24\sqrt{6t} \right) \leq 2e^{-t},$$

where

$$R_S \equiv 2^{-S} \sqrt{m} + 14\sqrt{6} \sum_{s=0}^S 2^{-s} \sqrt{H_\square(2^{-s}, \mathcal{F}_j, \|\cdot\|)} + \frac{36MH_\square(1, \mathcal{F}_j, \|\cdot\|)}{\sqrt{m}}.$$

Proof of Lemma 1.4:

The theorem we present here is slightly different from the original Theorem 8 in van der Geer and Lederer (2013) in that we have used the generalized bracketing entropy in Definition A.2 in the manuscript. Although the original version is presented with the usual L_2 -bracketing entropy for univariate functions, the whole “chaining along a tree” argument will still go through, if we replace the $\|\cdot\|$ and $|\cdot|_q$ norms and bracketing entropies in their proofs by our generalized versions to multivariate functions as in Definition A.2. \square

Lemma 1.5 (Generalization of Wong and Shen (1995) Theorem 1) *Assume (A3) holds. Then for any $\delta > 0$, there exist positive constants q_1, q_2 that depend on D_1, D_2 , such that for all subsets $Y_{[j]}$ with $j = 1, \dots, k$ and all sufficiently large m ,*

$$P_{\theta_0}^{(n)} \left(\sup_{h_{mj}(\theta, \theta_0) \geq \delta} \prod_{i=1}^m \frac{p(Y_{ji} | \theta)}{p(Y_{ji} | \theta_0)} \geq \exp(-q_1 m \delta^2) \right) \leq 4 \exp(-q_2 m \delta^2) \quad (34)$$

Proof of Lemma 1.5:

We consider the class of functions

$$\hat{\tilde{z}}_j(r) = \left\{ \frac{\tilde{z}_j(\theta)}{2\sqrt{2c_{2\tau}r}} : \theta \in \Theta \text{ satisfies } h_{mj}(\theta, \theta_0) \leq r \right\},$$

for a fixed $r > 0$. This class is a rescaled version of

$$\tilde{z}_j(\{\theta \in \Theta : h_{mj}(\theta, \theta_0) \leq r\}) = \left\{ \tilde{z}_j(\theta) : \theta \in \Theta \text{ satisfies } h_{mj}(\theta, \theta_0) \leq r \right\}$$

as in Lemma 1.3. By Lemma 1.2, $\hat{\tilde{z}}_j(r)$ satisfies Conditions (i) and (ii) in Lemma 1.4 with the constant $M = 1/(\sqrt{2c_{2\tau}r})$. Therefore the concentration inequality in Lemma 1.4 holds for $\hat{\tilde{z}}_j(r)$.

We first simplify the term R_S in the inequality. R_S involves the L_2 -bracketing entropy of the class $\hat{\tilde{z}}_j(r)$. Since $H_{[]}(\mathbf{u}, \hat{\tilde{z}}_j(r), \|\cdot\|)$ is nonincreasing in \mathbf{u} , we have

$$\begin{aligned} \sum_{s=0}^S 2^{-s} \sqrt{H_{[]} \left(2^{-s}, \hat{\tilde{z}}_j(r), \|\cdot\| \right)} &\leq 2 \sum_{s=0}^S \int_{2^{-(s+1)}}^{2^{-s}} \sqrt{H_{[]} \left(\mathbf{u}, \hat{\tilde{z}}_j(r), \|\cdot\| \right)} d\mathbf{u} \\ &= 2 \int_{2^{-(S+1)}}^1 \sqrt{H_{[]} \left(\mathbf{u}, \hat{\tilde{z}}_j(r), \|\cdot\| \right)} d\mathbf{u} \\ &\stackrel{(i)}{=} 2 \int_{2^{-(S+1)}}^1 \sqrt{H_{[]} \left(2\sqrt{2c_{2\tau}r}\mathbf{u}, \tilde{z}_j(\{\theta \in \Theta : h_{mj}(\theta, \theta_0) \leq r\}), \|\cdot\| \right)} d\mathbf{u} \\ &= \frac{1}{\sqrt{2c_{2\tau}r}} \int_{2^{-S}\sqrt{2c_{2\tau}r}}^{2\sqrt{2c_{2\tau}r}} \sqrt{H_{[]} \left(\mathbf{u}, \tilde{z}_j(\{\theta \in \Theta : h_{mj}(\theta, \theta_0) \leq r\}), \|\cdot\| \right)} d\mathbf{u} \\ &\stackrel{(ii)}{\leq} \frac{1}{\sqrt{2c_{2\tau}r}} \int_{2^{-S}\sqrt{2c_{2\tau}r}}^{2\sqrt{2c_{2\tau}r}} \sqrt{H_{[]} \left(\mathbf{u}/(2e^{\tau/2}), \mathcal{P}_j(\{\theta \in \Theta : h_{mj}(\theta, \theta_0) \leq r\}), h_{mj} \right)} d\mathbf{u} \\ &= \frac{\sqrt{2e^{\tau}}}{\sqrt{c_{2\tau}r}} \int_{2^{-(S+1)}\sqrt{2c_{2\tau}e^{-\tau}r}}^{\sqrt{2c_{2\tau}e^{-\tau}r}} \sqrt{H_{[]} \left(\mathbf{u}, \mathcal{P}_j(\{\theta \in \Theta : h_{mj}(\theta, \theta_0) \leq r\}), h_{mj} \right)} d\mathbf{u}, \end{aligned} \quad (35)$$

where (i) follows from the scaling relation between $\hat{\tilde{z}}_j(r)$ and $\tilde{z}_j(\{\theta \in \Theta : h_{mj}(\theta, \theta_0) \leq r\})$, and (ii) follows from Lemma 1.3. We choose integer $S \geq 1$ such that $2^{-(S+2)} \leq \sqrt{2c_{2\tau}e^{-\tau}r}/2^{1/2} \leq 2^{-(S+1)}$. It is possible to do so because we only need to consider $r \leq \sqrt{2}$ (since the h_{mj} distance is upper bounded by $\sqrt{2}$), $c_{2\tau}e^{-\tau} \leq 1/2$ for all $\tau \geq 0$, and it is guaranteed that $\sqrt{2c_{2\tau}e^{-\tau}r}/2^{1/2} \leq \sqrt{2}/2^{1/2} < 1/4$.

Now we can apply (A3) to (35) and obtain that uniformly over all $j = 1, \dots, k$,

$$\begin{aligned} \sum_{s=0}^S 2^{-s} \sqrt{H_{[]} \left(2^{-s}, \hat{\tilde{z}}_j(r), \|\cdot\| \right)} &\leq \frac{\sqrt{2e^{\tau}}}{\sqrt{c_{2\tau}r}} \int_{(\sqrt{2c_{2\tau}e^{-\tau}r})^{2/2^{1/2}}}^{\sqrt{2c_{2\tau}e^{-\tau}r}} \sqrt{\Psi(\mathbf{u}, r)} d\mathbf{u} \\ &\leq \frac{\sqrt{2e^{\tau}}}{\sqrt{c_{2\tau}r}} \cdot D_2 \sqrt{m} \left(\frac{\sqrt{2c_{2\tau}e^{-\tau}r}}{D_1} \right)^2 = \frac{2D_2\sqrt{2c_{2\tau}e^{-\tau}}}{D_1^2} \sqrt{mr}. \end{aligned} \quad (36)$$

Furthermore, since (A3) says $\Psi(\mathbf{u}, r)$ is nonincreasing in \mathbf{u} , we can also derive that

$$\begin{aligned} H_{[]} \left(1, \hat{\tilde{z}}_j(r), \|\cdot\| \right) &= H_{[]} \left(2\sqrt{2c_{2\tau}r}, \tilde{z}_j(\{\theta \in \Theta : h_{mj}(\theta, \theta_0) \leq r\}), \|\cdot\| \right) \\ &\leq H_{[]} \left(\sqrt{2c_{2\tau}e^{-\tau}r}, \mathcal{P}_j(\{\theta \in \Theta : h_{mj}(\theta, \theta_0) \leq r\}), h_{mj} \right) \leq \Psi \left(\sqrt{2c_{2\tau}e^{-\tau}r}, r \right) \end{aligned}$$

$$\begin{aligned}
&\leq \left[\frac{1}{\sqrt{2c_{2\tau}e^{-\tau}r} - (\sqrt{2c_{2\tau}e^{-\tau}r})^2/2^{12}} \int_{(\sqrt{2c_{2\tau}e^{-\tau}r})^2/2^{12}}^{\sqrt{2c_{2\tau}e^{-\tau}r}} \sqrt{\Psi(u, r)} du \right]^2 \\
&\leq \frac{2D_2^2 c_{2\tau} e^{-\tau} m r^2}{D_1^4 (1 - \sqrt{2c_{2\tau}e^{-\tau}r}/2^{12})^2} \leq \frac{8D_2^2 c_{2\tau} e^{-\tau} m r^2}{D_1^4},
\end{aligned} \tag{37}$$

where in the last step, we used the fact that $\sqrt{2c_{2\tau}e^{-\tau}r}/2^{12} < 1/2$.

By our choice $2^{-(S+2)} \leq \sqrt{2c_{2\tau}e^{-\tau}r}/2^{12}$, it follows from (36) and (37) that

$$\begin{aligned}
\min_{S \in \mathbb{N}} R_S &\leq \frac{\sqrt{2c_{2\tau}e^{-\tau}}}{2^{10}} \cdot \sqrt{mr} + 14\sqrt{6} \cdot \frac{2D_2\sqrt{2c_{2\tau}e^{-\tau}}}{D_1^2} \sqrt{mr} + \frac{36}{\sqrt{2c_{2\tau}mr}} \cdot \frac{8D_2^2 c_{2\tau} e^{-\tau} m r^2}{D_1^4} \\
&\leq \left[\frac{\sqrt{2}}{2^{10}} + 56\sqrt{3} \frac{D_2}{D_1^2} + 144\sqrt{2} \frac{D_2^2}{D_1^4} e^{-\tau/2} \right] \sqrt{c_{2\tau}e^{-\tau}} \sqrt{mr}.
\end{aligned} \tag{38}$$

In the inequality of Lemma 1.4, we let $t = c_{3\tau} m r^2$ with integer $\ell \geq 1$ and constant $c_{3\tau}$ to be chosen later, then with probability at least $1 - 2e^{-c_{3\tau} m r^2}$, the empirical process

$$\sup_{f \in \hat{\mathcal{Z}}_j(r)} \frac{1}{\sqrt{m}} \sum_{i=1}^m \left[f_i(Y_{ji}) - \mathbb{E}_{P_{\theta_0}} f_i(Y_{ji}) \right]$$

will not exceed the following upper bound

$$\begin{aligned}
&\min_{S \in \mathbb{N}} R_S + \frac{36M(1+t)}{\sqrt{m}} + 24\sqrt{6t} \\
&\leq \left[\frac{\sqrt{2}}{2^{10}} + 56\sqrt{3} \frac{D_2}{D_1^2} + 144\sqrt{2} \frac{D_2^2}{D_1^4} e^{-\tau/2} \right] \sqrt{c_{2\tau}e^{-\tau}} \sqrt{mr} + \frac{36(1+c_{3\tau}mr^2)}{\sqrt{2c_{2\tau}mr}} + 24\sqrt{6c_{3\tau}mr} \\
&= c_{4\tau} \sqrt{mr} + \frac{36}{\sqrt{2c_{2\tau}mr}},
\end{aligned}$$

where $c_{4\tau} = \left[\frac{\sqrt{2}}{2^{10}} + 56\sqrt{3} \frac{D_2}{D_1^2} + 144\sqrt{2} \frac{D_2^2}{D_1^4} e^{-\tau/2} \right] \sqrt{c_{2\tau}e^{-\tau}} + \frac{36c_{3\tau}}{\sqrt{2c_{2\tau}}} + 24\sqrt{6c_{3\tau}}$. As a result, the empirical process on the class $\tilde{\mathcal{Z}}_j(\{\theta \in \Theta : h_{mj}(\theta, \theta_0) \leq r\})$ satisfies that with probability at least $1 - 2e^{-c_{3\tau} m r^2}$,

$$\begin{aligned}
&\sup_{\theta \in \Theta : h_{mj}(\theta, \theta_0) \leq r} \frac{1}{\sqrt{m}} \sum_{i=1}^m \left[\tilde{Z}_{ji}(\theta) - \mathbb{E}_{P_{\theta_0}} \tilde{Z}_{ji}(\theta) \right] \leq 2\sqrt{2c_{2\tau}r} \cdot \left(c_{4\tau} \sqrt{mr} + \frac{36}{\sqrt{2c_{2\tau}mr}} \right) \\
&\leq 2\sqrt{2c_{2\tau}c_{4\tau}} \cdot \sqrt{mr^2} + \frac{72}{\sqrt{m}}.
\end{aligned} \tag{39}$$

On the set $\{\theta \in \Theta : r \leq h_{mj}(\theta, \theta_0) \leq 2r\}$, we have $h_{mj}^2(\theta, \theta_0) \geq r^2$. By Lemma 1.1, it follows that

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E} \tilde{Z}_{ji}(\theta) \leq -(1 - c_{1\tau})r^2.$$

This together with (39) implies that with probability at least $1 - 2e^{-c_{3\tau} m r^2}$,

$$\sup_{\theta \in \Theta : r \leq h_{mj}(\theta, \theta_0) \leq 2r} \frac{1}{m} \sum_{i=1}^m \tilde{Z}_{ji}(\theta) \leq -(1 - c_{1\tau} - 8\sqrt{2c_{2\tau}c_{4\tau}})r^2 + \frac{72}{m}.$$

Now we choose τ such that $e^{-\tau/2} = 1/32$. Then $c_{1\tau} < 0.07$, $29 < c_{2\tau} < 30$. Set $c_{3\tau} = 1/2^{30}$. By (A3), $D_2 \leq D_1^2/2^{12}$. So

$$8\sqrt{2c_{2\tau}c_{4\tau}} \leq 8\sqrt{60} \cdot \left\{ \left[\frac{\sqrt{2}}{2^{10}} + \frac{56\sqrt{3}}{2^{12}} + \frac{144\sqrt{2}}{16 \cdot 2^{24}} \right] \sqrt{\frac{30}{2^{10}}} + \frac{36}{\sqrt{58} \cdot 2^{30}} + 24\sqrt{\frac{6}{2^{30}}} \right\} < 0.377,$$

which implies that $1 - c_{1\tau} - 8\sqrt{2c_{2\tau}c_{4\tau}} > 0.55$. Thus we have proved that for any $0 < r \leq \sqrt{2}$, uniformly for all $j = 1, \dots, k$ and for all sufficiently large m ,

$$\begin{aligned} & P_{\theta_0}^{(n)} \left(\sup_{\{\theta \in \Theta: r \leq h_{mj}(\theta, \theta_0) \leq 2r\}} \frac{1}{m} \sum_{i=1}^m Z_{ji}(\theta) \geq -0.55r^2 + \frac{72}{m} \right) \\ & \leq P_{\theta_0}^{(n)} \left(\sup_{\{\theta \in \Theta: r \leq h_{mj}(\theta, \theta_0) \leq 2r\}} \frac{1}{m} \sum_{i=1}^m \tilde{Z}_{ji}(\theta) \geq -0.55r^2 + \frac{72}{m} \right) \leq 2 \exp(-mr^2/2^{30}). \end{aligned} \quad (40)$$

Finally for a given δ , we set $r = 2^\ell \delta$ for integers $\ell \geq 0$, $m > 72/0.05/\delta = 1440/\delta$, and let L be the smallest integer such that $2^L \delta^2 > 2$. It follows that

$$\begin{aligned} & P_{\theta_0}^{(n)} \left(\sup_{\{\theta \in \Theta: h_{mj}(\theta, \theta_0) \geq \delta\}} \prod_{i=1}^m \frac{p(Y_{ji} | \theta)}{p(Y_{ji} | \theta_0)} \geq \exp(-0.5m\delta^2) \right) \\ & \leq P_{\theta_0}^{(n)} \left(\sup_{\{\theta \in \Theta: h_{mj}(\theta, \theta_0) \geq \delta\}} \frac{1}{m} \sum_{i=1}^m Z_{ji}(\theta) \geq -0.55\delta^2 + \frac{72}{m} \right) \\ & \leq \sum_{\ell=0}^L P_{\theta_0}^{(n)} \left(\sup_{\{\theta \in \Theta: 2^\ell \delta \leq h_{mj}(\theta, \theta_0) \leq 2^{\ell+1} \delta\}} \frac{1}{m} \sum_{i=1}^m Z_{ji}(\theta) \geq -0.55(2^\ell \delta)^2 + \frac{72}{m} \right) \\ & \leq \sum_{\ell=0}^L 2 \exp(-2^{2\ell} m \delta^2 / 2^{30}) \leq 4 \exp(-m \delta^2 / 2^{30}). \end{aligned}$$

We set $q_1 = 0.5$, $q_2 = 1/2^{30}$ and complete the proof. \square

Lemma 1.6 Assume (A3) holds. Then for any $\delta > 0$, there exist positive constants r_1, r_2 that depend on g_1, g_2, κ, c_π , such that for every subset $Y_{[j]}$ ($j = 1, \dots, k$), for any $t \geq \epsilon_m^{2\alpha}$,

$$P_{\theta_0}^{(n)} \left(\int_{\Theta} \prod_{i=1}^m \frac{p(Y_{ji}|\theta)^\gamma}{p(Y_{ji}|\theta_0)^\gamma} \Pi(d\theta) \leq \exp(-r_1 n t) \right) \leq \exp(-r_2 m t).$$

Proof of Lemma 1.6:

Define the event Θ_{ϵ_m} as

$$\Theta_{\epsilon_m} = \left\{ \theta \in \Theta : \frac{1}{m} \sum_{i=1}^m E_{P_{\theta_0}} \exp \left(\log_+ \frac{p(Y_{ji}|\theta_0)}{p(Y_{ji}|\theta)} \right) - 1 \leq \epsilon_m^{2\alpha} \right\}$$

then (A4) can be written as $\Pi(\Theta_{\epsilon_m}) \geq \exp(-c_\pi n \epsilon_m^{2\alpha})$. For $A \subseteq \Theta$, let $\Pi_{\epsilon_m}(A) = \Pi(A \cap \Theta_{\epsilon_m}) / \Pi(\Theta_{\epsilon_m})$ be the prior measure Π restricted measure to the set Θ_{ϵ_m} . For the left-hand side of the conclusion, we have

$$P_{\theta_0}^{(n)} \left(\int_{\Theta} \prod_{i=1}^m \frac{p(Y_{ji}|\theta)^\gamma}{p(Y_{ji}|\theta_0)^\gamma} \Pi(d\theta) \leq \exp(-r_1 n t) \right)$$

$$\begin{aligned}
& \stackrel{(i)}{\leq} P_{\theta_0}^{(n)} \left(\int_{\Theta_{\epsilon_m}} \prod_{i=1}^m \frac{p(Y_{ji}|\theta)^\gamma}{p(Y_{ji}|\theta_0)^\gamma} \Pi(d\theta) \leq \exp(-r_1 n t) \right) \\
& \leq P_{\theta_0}^{(n)} \left(\Pi(\Theta_{\epsilon_m}) \cdot \int_{\Theta_{\epsilon_m}} \prod_{i=1}^m \frac{p(Y_{ji}|\theta)^\gamma}{p(Y_{ji}|\theta_0)^\gamma} \Pi_{\epsilon_m}(d\theta) \leq \exp(-r_1 n t) \right) \\
& \stackrel{(ii)}{\leq} P_{\theta_0}^{(n)} \left(\int_{\Theta_{\epsilon_m}} \prod_{i=1}^m \frac{p(Y_{ji}|\theta)^\gamma}{p(Y_{ji}|\theta_0)^\gamma} \Pi_{\epsilon_m}(d\theta) \leq \exp(-r_1 n t + c_\pi n \epsilon_m^{2\alpha}) \right) \\
& \stackrel{(iii)}{\leq} P_{\theta_0}^{(n)} \left(\sum_{i=1}^m \int_{\Theta_{\epsilon_m}} \log \frac{p(Y_{ji}|\theta_0)}{p(Y_{ji}|\theta)} \Pi_{\epsilon_m}(d\theta) \geq r_1 g_1 m t - c_\pi g_2 m \epsilon_m^{2\alpha} \right) \\
& \stackrel{(iv)}{\leq} P_{\theta_0}^{(n)} \left(\sum_{i=1}^m \left[\int_{\Theta_{\epsilon_m}} \log \frac{p(Y_{ji}|\theta_0)}{p(Y_{ji}|\theta)} \Pi_{\epsilon_m}(d\theta) - E_{P_{\theta_0}} \int_{\Theta_{\epsilon_m}} \log \frac{p(Y_{ji}|\theta_0)}{p(Y_{ji}|\theta)} \Pi_{\epsilon_m}(d\theta) \right] \right. \\
& \quad \left. \geq r_1 g_1 m t - (c_\pi g_2 + \kappa^{-1}) m \epsilon_m^{2\alpha} \right). \tag{41}
\end{aligned}$$

In the derivation above, (i) follows from making the region of the integral smaller; (ii) is from (A4); (iii) is an application of Jensen's inequality and uses the fact that $g_1 \gamma m \leq n \leq g_2 \gamma m$; (iv) follows by using Fubini's theorem, the inequality $x \leq (e^{\kappa x} - 1)/\kappa$ for $x \geq 0$, and (A4):

$$\begin{aligned}
& \sum_{i=1}^m E_{P_{\theta_0}} \int_{\Theta_{\epsilon_m}} \log \frac{p(Y_{ji}|\theta_0)}{p(Y_{ji}|\theta)} \Pi_{\epsilon_m}(d\theta) \leq \int_{\Theta_{\epsilon_m}} \sum_{i=1}^m E_{P_{\theta_0}} \log_+ \frac{p(Y_{ji}|\theta_0)}{p(Y_{ji}|\theta)} \Pi_{\epsilon_m}(d\theta) \\
& \leq \int_{\Theta_{\epsilon_m}} \frac{1}{\kappa} \sum_{i=1}^m \left[E_{P_{\theta_0}} \exp \left(\kappa \log_+ \frac{p(Y_{ji}|\theta_0)}{p(Y_{ji}|\theta)} \right) - 1 \right] \Pi_{\epsilon_m}(d\theta) \leq \kappa^{-1} m \epsilon_m^{2\alpha}.
\end{aligned}$$

Now in (41), if we choose r_1 large enough and let $Z_i = \int_{\Theta_{\epsilon_m}} \log \frac{p(Y_{ji}|\theta_0)}{p(Y_{ji}|\theta)} \Pi_{\epsilon_m}(d\theta)$, we can use the Bernstein's inequality (Corollary 2.10 in Massart (2003)) to control the large deviation for the sum of Z_i 's. By inspecting the conditions for this inequality, we can see that for any integer $\ell \geq 2$, by the inequality $(\kappa x)^\ell / \ell! \leq e^{\kappa x} - 1$, (A4), and the Jensen's inequality,

$$\sum_{i=1}^m E_{P_{\theta_0}} [(Z_i)_+^\ell] \leq \frac{\ell!}{\kappa^\ell} \sum_{i=1}^m \left[\exp \left(\kappa E_{P_{\theta_0}}(Z_i)_+ \right) - 1 \right] \leq \frac{\ell!}{\kappa^\ell} m \epsilon_m^{2\alpha}.$$

Therefore in the Corollary 2.10, we can set $c = \kappa^{-1}$, $v = 2\kappa^{-2} m \epsilon_m^{2\alpha}$, $x = r_1 g_1 m t - (c_\pi g_2 + \kappa^{-1}) m \epsilon_m^{2\alpha}$. If we choose $r_1 > (c_\pi g_2 + 3\kappa^{-1})/g_1$, then since $t \geq \epsilon_m^{2\alpha}$, we have $x > 2\kappa^{-1} m t \geq 2\kappa^{-1} m \epsilon_m^{2\alpha}$. Hence it follows that $v/c = 2\kappa^{-1} m \epsilon_m^{2\alpha} < x$. We apply the Bernstein's inequality to (41) and obtain that

$$\begin{aligned}
& P_{\theta_0}^{(n)} \left(\sum_{i=1}^m [Z_i - E_{P_{\theta_0}} Z_i] \geq r_1 g_1 m t - (c_\pi g_2 + \kappa^{-1}) m \epsilon_m^{2\alpha} \right) \\
& \leq \exp \left[-\frac{x^2}{2(v + cx)} \right] \leq \exp \left(-\frac{x}{4c} \right) \leq \exp \left(-\frac{\kappa x}{4} \right) \leq \exp \left(-\frac{m t}{2} \right).
\end{aligned}$$

We set $r_2 = 1/2$ and complete the proof. \square

Lemma 1.7 *Let $\bar{\nu}$ denote the W_2 barycenter of N measures ν_1, \dots, ν_N in $\mathcal{P}_2(\Theta)$. Then for any $\theta_0 \in \Theta$,*

$$W_2(\bar{\nu}, \delta_{\theta_0}) \leq \frac{1}{N} \sum_{j=1}^N W_2(\nu_j, \delta_{\theta_0}).$$

Proof of Lemma 1.7:

Theorem 4.1 in Agueh and Carlier (2011) shows that $\bar{\nu} = \bar{T} \# \nu_1$, where

$$\bar{T}(\theta) = \frac{1}{N} \sum_{j=1}^N T_j^1(\theta),$$

for any $\theta \in \Theta$ and T_j^1 is the optimal transport map that pushes ν_1 forward to ν_j , i.e. $\nu_j = T_j^1 \# \nu_1$ and T_1^1 is the identity operator. We use the property of the ρ metric in (A5) and obtain that

$$\begin{aligned} W_2^2(\bar{\nu}, \delta_{\theta_0}) &= \int_{\Theta} \rho^2 \left(\frac{1}{N} \sum_{j=1}^N T_j^1(\theta), \theta_0 \right) \nu_1(d\theta) \leq \int_{\Theta} \left[\frac{1}{N} \sum_{j=1}^N \rho(T_j^1(\theta), \theta_0) \right]^2 \nu_1(d\theta) \\ &= \frac{1}{N^2} \sum_{j=1}^N \int_{\Theta} \rho^2(T_j^1(\theta), \theta_0) \nu_1(d\theta) + \frac{1}{N^2} \sum_{j \neq l} \int_{\Theta} \rho(T_j^1(\theta), \theta_0) \rho(T_l^1(\theta), \theta_0) \nu_1(d\theta). \end{aligned} \quad (42)$$

For each term in the second summation, we apply the Cauchy-Schwartz inequality and have

$$\begin{aligned} &\int_{\Theta} \rho(T_j^1(\theta), \theta_0) \rho(T_l^1(\theta), \theta_0) \nu_1(d\theta) \\ &\leq \sqrt{\int_{\Theta} \rho^2(T_j^1(\theta), \theta_0) \nu_1(d\theta)} \cdot \sqrt{\int_{\Theta} \rho^2(T_l^1(\theta), \theta_0) \nu_1(d\theta)} = W_2(\nu_j, \delta_{\theta_0}) \cdot W_2(\nu_l, \delta_{\theta_0}). \end{aligned}$$

Therefore in (42),

$$\begin{aligned} W_2^2(\bar{\nu}, \delta_{\theta_0}) &\leq \frac{1}{N^2} \sum_{j=1}^N W_2^2(\nu_j, \delta_{\theta_0}) + \frac{1}{N^2} \sum_{j \neq l} W_2(\nu_j, \delta_{\theta_0}) \cdot W_2(\nu_l, \delta_{\theta_0}) \\ &= \left[\frac{1}{N} \sum_{j=1}^N W_2(\nu_j, \delta_{\theta_0}) \right]^2, \end{aligned}$$

and hence the conclusion follows. \square

2 Experiments

2.1 Simulated data: finite mixture of Gaussians

Consider the set of L mixture of Gaussians. If $\mathbf{y}_i \in \mathbb{R}^p$ is the i th observation ($i = 1, \dots, n$) sampled from a mixture of L Gaussians, then

$$p(\mathbf{y}_i | \theta) = \sum_{l=1}^L \pi_l \mathcal{N}_p(\mathbf{y} | \boldsymbol{\mu}_l, \Sigma_l), \quad (43)$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_L)$ lies in the $(L - 1)$ -simplex, $\boldsymbol{\mu}_l$ and Σ_l ($l = 1, \dots, L$) are the mean and covariance parameters of a p -variate Gaussian distribution, and $\theta = \{\boldsymbol{\pi}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_L, \Sigma_1, \dots, \Sigma_L\}$. We cluster $\mathbf{y}_1, \dots, \mathbf{y}_n$ into L clusters using K-Means clustering and randomly split the members of every cluster into k subsets. This ensures that full-data are split into k subsets such that the mixture proportions are represented

in every subset. Let $\mathbf{y}_{j1}, \dots, \mathbf{y}_{jm}$ represent the data on subset j ($j = 1, \dots, k$). The hierarchical model for the data on subset j is

$$\mathbf{y}_{ji} | z_{ji}, \theta \sim \mathcal{N}_p(\boldsymbol{\mu}_{z_{ji}}, \boldsymbol{\Sigma}_{z_{ji}}), \quad z_{ji} \sim \sum_{h=1}^L \pi_h \delta_h, \quad (44)$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(L^{-1}, \dots, L^{-1}), \quad \boldsymbol{\mu}_h | \boldsymbol{\Sigma}_h \sim \mathcal{N}_p(\mathbf{0}, 100\boldsymbol{\Sigma}_h), \quad \boldsymbol{\Sigma}_h \sim \text{Inverse-Wishart}(2, 4I_p), \quad h = 1, \dots, L,$$

where 2 is the prior degrees of freedom and $4I_p$ is the scale matrix of the Inverse-Wishart distribution.

The posterior distribution of θ after stochastic approximation is derived using standard arguments for finite mixture of Gaussians. The likelihood given $\mathbf{y}_{j1}, \dots, \mathbf{y}_{jm}$ and latent variables z_{j1}, \dots, z_{jm} is

$$L_j(\{\boldsymbol{\mu}_h\}_{h=1}^L, \{\boldsymbol{\Sigma}_h\}_{h=1}^L, \{\pi_h\}_{h=1}^L) = \prod_{h=1}^L (2\pi^p |\boldsymbol{\Sigma}_h|)^{-\frac{\#h_j}{2}} e^{-\frac{1}{2} \sum_{i=1}^m 1(z_{ji}=h) (\mathbf{y}_{ji} - \boldsymbol{\mu}_h)^T \boldsymbol{\Sigma}_h^{-1} (\mathbf{y}_{ji} - \boldsymbol{\mu}_h)} \pi_h^{\#h_j}, \quad (45)$$

where $1(z_{ji} = h)$ is 1 if $z_{ji} = h$ and 0 otherwise and $\#h_j = \sum_{i=1}^m 1(z_{ji} = h)$. For stochastic approximation, we raise L_j in (45) to the power γ and obtain

$$L_j^\gamma(\{\boldsymbol{\mu}_h\}_{h=1}^L, \{\boldsymbol{\Sigma}_h\}_{h=1}^L, \{\pi_h\}_{h=1}^L) = \prod_{h=1}^L (2\pi^p |\boldsymbol{\Sigma}_h|)^{-\frac{\gamma \#h_j}{2}} e^{-\frac{\gamma}{2} \sum_{i=1}^m 1(z_{ji}=h) (\mathbf{y}_{ji} - \boldsymbol{\mu}_h)^T \boldsymbol{\Sigma}_h^{-1} (\mathbf{y}_{ji} - \boldsymbol{\mu}_h)} \pi_h^{\gamma \#h_j}. \quad (46)$$

If we use L_j^γ as the likelihood in (43), then the prior for θ in (43) and simple extensions of standard arguments for finite mixture of Gaussians imply that the analytic form of full conditional densities of the parameters are as follows. Define

$$h_j = \{i : 1(z_{ji} = h) = 1, i = 1, \dots, m\}, \quad \bar{\mathbf{y}}_{h_j} = \frac{1}{\#h_j} \sum_{i \in h_j} \mathbf{y}_{ji} \quad (h = 1, \dots, L), \quad (47)$$

and the complete conditionals of the parameters are

$$\begin{aligned} \boldsymbol{\pi}_j | \text{rest} &\sim \text{Dirichlet} \left(\gamma \sum_{i=1}^m 1(z_{ji} = 1) + L^{-1}, \dots, \gamma \sum_{i=1}^m 1(z_{ji} = L) + L^{-1} \right) \\ \boldsymbol{\mu}_{jh} | \text{rest} &\sim \text{Normal} \left\{ \frac{\gamma \#h_j}{0.01 + \gamma \#h_j} \bar{\mathbf{y}}_{h_j}, \frac{1}{0.01 + \gamma \#h_j} \boldsymbol{\Sigma}_h \right\}, \\ \boldsymbol{\Sigma}_{jh} | \text{rest} &\sim \text{Inverse-Wishart} \left\{ \gamma \#h_j + 3, \sum_{i \in h_j} (\mathbf{y}_{ji} - \bar{\mathbf{y}}_{h_j})(\mathbf{y}_{ji} - \bar{\mathbf{y}}_{h_j})^T + \frac{0.01 \cdot \gamma \#h_j}{0.01 + \gamma \#h_j} \bar{\mathbf{y}}_{h_j} \bar{\mathbf{y}}_{h_j}^T + 4I_p \right\} \\ z_{ji} | \text{rest} &\sim \sum_{h=1}^L p_{jh} \delta_h, \quad p_{jh} = \frac{\pi_{jh} \mathcal{N}_p(\mathbf{y}_{ji} | \boldsymbol{\mu}_{jh}, \boldsymbol{\Sigma}_{jh})}{\sum_{\tilde{h}=1}^L \pi_{j\tilde{h}} \mathcal{N}_p(\mathbf{y}_{ji} | \boldsymbol{\mu}_{j\tilde{h}}, \boldsymbol{\Sigma}_{j\tilde{h}})} \end{aligned} \quad (48)$$

for $h = 1, \dots, L$ and $i = 1, \dots, m$; see Chapter 9 in Bishop (2006) for details.

All full conditionals are analytically tractable in terms of standard distributions. The Gibbs sampler iterates between the following four steps:

1. Sample $\boldsymbol{\pi}_j$ from Dirichlet distribution in (48).
2. Sample $\boldsymbol{\mu}_{jh}$ from Normal distribution in (48) for $h = 1, \dots, L$.
3. Sample $\boldsymbol{\Sigma}_{jh}$ from Inverse-Wishart distribution in (48) for $h = 1, \dots, L$.
4. Sample z_{ji} from categorical distribution in (48) for $i = 1, \dots, m$.

2.2 Simulated data analysis: Linear mixed effects model

The conditional densities of β and Σ in two steps. First, the sampling model of the linear mixed effects model implies that the likelihood of i th observation in j th subset is

$$L_{ji} = \int_{\mathbb{R}^q} \mathcal{N}_{n_i}(\mathbf{y}_{ji} | X_i \beta + Z_i \mathbf{u}_i, \tau^2 I_{n_i}) \mathcal{N}_q(\mathbf{u}_i | \mathbf{0}, \Sigma) d\mathbf{u}_i = \mathcal{N}_{n_i}(\mathbf{y}_{ji} | X_i \beta, Z_i \Sigma Z_i^T + \tau^2 I_{n_i}). \quad (49)$$

This implies that likelihood of β and Σ after stochastic approximation is

$$L_j^\gamma = \prod_{i=1}^m \{L_{ji}\}^\gamma = \prod_{i=1}^m \{\mathcal{N}_{n_i}(\mathbf{y}_{ji} | X_i \beta, Z_i \Sigma Z_i^T + \tau^2 I_{n_i})\}^\gamma. \quad (50)$$

Second, the j th subset posterior distribution of β and Σ after stochastic approximation is calculated using L_j^γ as the likelihood and the same priors for β and Σ as in the sampling model of the linear mixed effects model. Instead of finding the analytic form of the posterior density, we use the `increment_log_prob` function in Stan (Stan Development Team, 2014) to specify that the likelihood of \mathbf{y}_{ji} as $\{L_{ji}\}^\gamma$ (50) and use the default priors of β and Σ in Stan to obtain samples of (β, Σ) from the j th subset posterior after stochastic approximation. Table (10) describes the accuracies of CMC, SDP, VB, and WASP in approximating the full data marginal posterior distributions of fixed effects in the simulation.

Table 10: Accuracies of the approximate posteriors for β in linear mixed effects model simulation. The accuracies are averaged over 10 simulation replications and across all elements of β . Monte Carlo errors are in parenthesis. VB, variational Bayes; CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior.

VB	p = 4		p = 80	
	0.96 (0.01)		0.96 (0.01)	
	k = 10	k = 20	k = 10	k = 20
CMC	0.96 (0.01)	0.95 (0.02)	0.95 (0.02)	0.94 (0.03)
SDP	0.96 (0.01)	0.95 (0.02)	0.90 (0.06)	0.90 (0.06)
WASP	0.95 (0.01)	0.96 (0.02)	0.95 (0.02)	0.94 (0.03)

2.3 Simulated data analysis: Probablistic parafac model

The derivation of modified full conditional densities of unknown parameters involves one key modification in the Gibbs sampling algorithm of Dunson and Xing (2009). Assuming z_{ji} is given, the contribution of i th observation in j th subset to the likelihood after stochastic approximation is

$$L_{ji}^\gamma((\psi_h^{(1)})_{h=1}^{l^*}, \dots, (\psi_h^{(q)})_{h=1}^{l^*}, \dots, (\psi_h^{(p)})_{h=1}^{l^*}) = \left(\prod_{h=1}^{l^*} v_h \prod_{q=1}^p \prod_{l=1}^{d_q} \psi_{hl}^{(q)1(x_{jiq}=l, z_{ji}=h)} \right)^\gamma,$$

where $1(x_{jiq} = l, z_{ji} = h)$ is 1 if both conditions are true and is 0 otherwise and l^* is the maximum number of atoms in the stick breaking representation for the distribution of z_{ji} . The conditional posterior density of $\psi_h^{(q)}$ after stochastic approximation is proportional to

$$\prod_{l=1}^{d_q} \psi_{hl}^{(q)a_{jl}-1} \prod_{i=1}^m \prod_{l=1}^{d_q} \psi_{hl}^{(q)\gamma 1(x_{jiq}=l, z_{ji}=h)} = \prod_{l=1}^{d_q} \psi_{hl}^{(q)a_{jl}+\gamma \sum_{i=1}^m 1(x_{jiq}=l, z_{ji}=h)-1},$$

which implies that

$$\psi_{jh}^{(q)} \mid \text{rest} \sim \text{Dirichlet} (a_{q1} + \gamma 1(x_{jiq} = 1, z_{ji} = h), \dots, a_{qd_q} + \gamma 1(x_{jiq} = d_q, z_{ji} = h)) \quad (51)$$

for $q = 1, \dots, p$ and $h = 1, \dots, l^*$. The conditional densities of remaining parameters follow from Section 3.1 of Dunson and Xing (2009):

$$V_{jh} \mid \text{rest} \sim \text{Beta}(1 + \gamma \sum_{i=1}^m 1(z_{ji} = h), \alpha + \gamma \sum_{i=1}^m 1(z_{ji} > h)), \quad (52)$$

$$\alpha_j \mid \text{rest} \sim \text{Gamma}(a_\alpha + l^*, b_\alpha - \sum_{h=1}^{l^*} \log(1 - V_{jh})). \quad (53)$$

Finally, we update the posterior density of responsibility of every observation as

$$z_{ji} \mid \text{rest} \sim \sum_{h=1}^{l^*} p_{jh} \delta_h, \quad p_{jh} = \frac{v_{jh} \prod_{q=1}^p \psi_{hx_{jiq}}^{(q)}}{\sum_{h=1}^{l^*} v_{jh} \prod_{q=1}^p \psi_{hx_{jiq}}^{(q)}}, \quad (h = 1, \dots, l^*; i = 1, \dots, m), \quad (54)$$

where $v_{jh} = V_{jh} \prod_{l < h} (1 - V_{jl})$. The conditional posterior densities without stochastic approximation are obtained by substituting $\gamma = 1$ and $m = n$ in the full conditionals (51) – (53).

All full conditionals are analytically tractable in terms of standard distributions. The Gibbs sampler iterates between the following four steps:

1. Sample $\psi_{jh}^{(q)}$ from Dirichlet distribution in (51) for $q = 1, \dots, p$ and $h = 1, \dots, l^*$.
2. Sample V_{jh} from Beta distribution in (52) for $h = 1, \dots, l^*$.
3. Sample α_j from Gamma distribution in (53).
4. Sample z_{ji} from categorical distribution in (54) for $i = 1, \dots, m$.

We fix $a_{ql} = 1/d_q$ for $q = 1, \dots, p$ and $l = 1, \dots, d_q$.

2.4 Real data analysis: MovieLens data

Table 11 describes the accuracies of CMC, SDP, VB, and WASP in approximating the full data marginal posterior distributions of fixed effects in the MovieLens data analysis.

Table 11: Accuracies of the approximate posteriors of the fixed effects in the linear mixed effects model for MovieLens data. The accuracies are averaged over 10 replications. Monte Carlo errors are in parenthesis. CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior.

	β_{Action}	$\beta_{\text{Children} - \text{Action}}$	$\beta_{\text{Comedy} - \text{Action}}$	$\beta_{\text{Drama} - \text{Action}}$	$\beta_{\text{Popularity}}$	β_{Previous}
CMC	0.95 (0.02)	0.93 (0.02)	0.92 (0.02)	0.95 (0.02)	0.94 (0.02)	0.95 (0.03)
SDP	0.93 (0.03)	0.92 (0.03)	0.92 (0.04)	0.93 (0.04)	0.92 (0.04)	0.94 (0.02)
VB	0.00 (0.00)	0.34 (0.05)	0.64 (0.05)	0.15 (0.02)	0.01 (0.00)	0.00 (0.00)
WASP	0.96 (0.01)	0.95 (0.01)	0.95 (0.02)	0.96 (0.01)	0.96 (0.01)	0.96 (0.01)

3 MATLAB Code for solving the WASP linear program

For ease of illustration, we consider the problem of estimating an atomic approximation of the WASP using three subset posterior distributions consisting of 50, 75, and 100 posterior samples from Gaussian distributions with means $\mu_1 = (3, 2)^T$, $\mu_2 = (2, 3)^T$, and $\mu_3 = (3, 3)^T$ and covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma$, where $\sigma_{11} = 1$, $\sigma_{22} = 3$, and $\sigma_{21} = 1.5$. The MATLAB function `WASP`, which estimates the atomic approximation of the WASP, requires two inputs. First, the support of the WASP. Second, the distance matrix between the atoms of the WASP and every subset posterior distribution. We assume that the WASP is supported on a grid of atoms estimated from all subset posterior samples, where mesh-size of the grid is specified by `grdsize`. We use the grid of atoms of the WASP to obtain the three distance matrices between the atoms of the WASP and every subset posterior distribution. The MATLAB function `WASP` uses the grid of atoms and distance matrices as inputs and estimates the weights of every atom in the support of the WASP. The empirical measure obtained using the atoms and their estimated weights is an atomic approximation of the true analytically intractable WASP.

```

rand('seed',0);

% Size of the grid
grdsize = 50;

% var-covar matrix for Multivariate Normal Distribution
sig = [1 1.5; 1.5 3];

% means for 3 subset posteriors
mu1 = [3 2];
mu2 = [2 3];
mu3 = [3 3];

% atoms for 3 subset posteriors; they only differ in their means.
spost{1} = mvnrnd(mu1, sig, 50);
spost{2} = mvnrnd(mu2, sig, 75);
spost{3} = mvnrnd(mu3, sig, 100);

% atoms for the WASP by forming a grid
lbd1 = min(cellfun(@(x) x(1), cellfun(@(x) min(x), spost,'UniformOutput', false)));
lbd2 = min(cellfun(@(x) x(2), cellfun(@(x) min(x), spost,'UniformOutput', false)));
ubd1 = max(cellfun(@(x) x(1), cellfun(@(x) max(x), spost,'UniformOutput', false)));
ubd2 = max(cellfun(@(x) x(2), cellfun(@(x) max(x), spost,'UniformOutput', false)));
[opostx, oposty] = meshgrid(linspace(lbd1, ubd1, grdsize), linspace(lbd2, ubd2, grdsize));
opost = [opostx(:) oposty(:)];

% calculate the pair-wise sq. euclidean distance between the atoms of subset
% posteriors and BarPost atoms
m11 = diag(spost{1} * spost{1}');
m22 = diag(spost{2} * spost{2}');
m33 = diag(spost{3} * spost{3}');
m00 = diag(opost * opost');

m01 = opost * spost{1}';
m02 = opost * spost{2}';
m03 = opost * spost{3}';

% calculate distance between atoms
d01 = bsxfun(@plus, bsxfun(@plus, -2 * m01, m11'), m00);
d02 = bsxfun(@plus, bsxfun(@plus, -2 * m02, m22'), m00);
d03 = bsxfun(@plus, bsxfun(@plus, -2 * m03, m33'), m00);

% initialize the wts b_1 ... b_K for subset posteriors atoms; see Eq. 21
b = cellfun(@(x) ones(size(x, 1), 1) / size(x, 1), spost, 'UniformOutput', false);

colMeasure = b;
distMat{1} = d01;
distMat{2} = d02;
distMat{3} = d03;

tic
[optSol, optObj, exitflag, output, lambda] = WASP(colMeasure, distMat);
toc

function [optSol, obj, exitflag, output, lambda] = WASP(colMeasure, distMat)
%% WASP function to calculate the Wasserstein Barycenter of a list of
% empirical measures and pairwise distance between atoms.
% Input:
% =====
% colMeasure:
% cell containing the wts of atoms of empirical measures.
%
% distMat:
% cell containing pairwise distance between atoms of empirical measures and

```

```

% the Wasserstein barycenter.
%
% Output:
% =====
% output of solving the LP (31) in the manuscript using 'linprog'

nsubset = length(distMat);
vecColMeasure = cell2mat(cellfun(@(x) x', colMeasure, 'UniformOutput', false));
nsample = size(vecColMeasure, 1);

vecDistMat = cell2mat(distMat);
vecDistMat = vecDistMat(:);

fmatCell = cellfun(@(x) kron(ones(1, size(x, 1)), eye(nsample)), colMeasure, 'UniformOutput', false);
fmat = blkdiag(fmatCell{:}); % F

hmatCell = cellfun(@(x) kron(eye(size(x, 1)), ones(1, nsample)), colMeasure, 'UniformOutput', false);
hmat = blkdiag(hmatCell{:}); % H

gmat = kron(ones(nsubset, 1), eye(nsample)); % G

% Aeq in the matlab linprog function
aMat = [zeros(1, nsample^2) ones(1, nsample);
        fmat          -gmat;
        hmat          zeros(nsample, nsample)
        1];
% beq in the matlab linprog function
bVec = [1;
        zeros(nsubset * nsample, 1);
        vecColMeasure
        1];
% f in the matlab linprog function
costVec = [vecDistMat;
           zeros(nsample, 1)];

% upper bds = 1 and lower bds = 0
lbd = zeros(nsample^2 + nsample, 1);
ubd = ones(nsample^2 + nsample, 1);

[optSol, obj, exitflag, output, lambda] = linprog(costVec, [], [], aMat, bVec, lbd, ubd);

```

References

- Agueh, M. and G. Carlier (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis* 43(2), 904–924.
- Ahn, S., A. Korattikara, and M. Welling (2012). Bayesian posterior sampling via stochastic gradient Fisher scoring. *Proceedings of the 29th International Conference on Machine Learning*.
- Alquier, P., N. Friel, R. Everitt, and A. Boland (2016). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing* 26(1-2), 29–47.
- Anderes, E., S. Borgwardt, and J. Miller (2016). Discrete Wasserstein barycenters: optimal transport for discrete data. *Mathematical Methods of Operations Research* 84(2), 389–409.
- Bardenet, R., A. Doucet, and C. Holmes (2015). On Markov chain Monte Carlo methods for tall data. *arXiv preprint arXiv:1505.02827*.
- Bickel, P. J. and D. A. Freedman (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics* 9(6), 1196–1217.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*, Volume 4. Springer New York.
- Broderick, T., N. Boyd, A. Wibisono, A. C. Wilson, and M. Jordan (2013). Streaming variational Bayes. In *Advances in Neural Information Processing Systems*, pp. 1727–1735.
- Carlier, G., A. Oberman, and E. Oudet (2015). Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis* 49(6), 1621–1642.

- Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300.
- Cuturi, M. and A. Doucet (2014). Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning, JMLR W&CP*, Volume 32.
- Dunson, D. B. and C. Xing (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* 104(487), 1042–1051.
- Faes, C., J. T. Ormerod, and M. P. Wand (2012). Variational Bayesian inference for parametric and nonparametric regression with missing data. *Journal of the American Statistical Association* 106(495), 959–971.
- Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* 97(458), 611–631.
- Gelman, A., A. Vehtari, P. Jylänki, C. Robert, N. Chopin, and J. P. Cunningham (2014). Expectation propagation as a way of life. *arXiv preprint arXiv:1412.4869*.
- Ghosal, S., J. K. Ghosh, and T. Samanta (1995). On convergence of posterior distributions. *The Annals of Statistics* 23(6), 2145–2152.
- Ghosal, S., J. K. Ghosh, and A. W. van Der Vaart (2000). Convergence rates of posterior distributions. *Annals of Statistics* 28(2), 500–531.
- Ghosal, S. and A. van Der Vaart (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics* 35(1), 192–223.
- Gurobi Optimization Inc. (2014). *Gurobi Optimizer Reference Manual Version 6.0.0*.
- Hoffman, M. D., D. M. Blei, C. Wang, and J. Paisley (2013). Stochastic variational inference. *Journal of Machine Learning Research* 14, 1303–1347.
- Ibragimov, I. A. and R. Z. Has’minskii (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.
- Johndrow, J. E., J. C. Mattingly, S. Mukherjee, and D. B. Dunson (2015). Approximations of Markov chains and High-Dimensional Bayesian Inference. *arXiv preprint arXiv:1508.03387v1*.
- Korattikara, A., Y. Chen, and M. Welling (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 181–189.
- Lan, S., B. Zhou, and B. Shahbaba (2014). Spherical Hamiltonian Monte Carlo for constrained target distributions. In *JMLR workshop and conference proceedings*, Volume 32, pp. 629. NIH Public Access.
- Lee, C. Y. Y. and M. P. Wand (2016). Streamlined mean field variational Bayes for longitudinal and multi-level data analysis. *Biometrical Journal* 58(4), 868–895.
- Maclaurin, D. and R. P. Adams (2015). Firefly Monte Carlo: Exact MCMC with Subsets of Data. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
- Massart, P. (2003). Concentration Inequalities and Model Selection. In *Ecole d’Eté de Probabilités de Saint-Flour XXXIII*, pp. 25–26. Springer-Verlag, Berlin Heidelberg.
- Minsker, S., S. Srivastava, L. Lin, and D. B. Dunson (2014). Robust and scalable Bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*.

- Miroshnikov, A. and E. Conlon (2014). *parallelMCMCcombine: Methods for combining independent subset Markov chain Monte Carlo posterior samples to estimate a posterior density given the full data set*. R package version 1.0.
- Neiswanger, W., C. Wang, and E. Xing (2014). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the 30th International Conference on Uncertainty in Artificial Intelligence*, pp. 623–632.
- Perry, P. O. (2016). Fast moment-based estimation for hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* forthcoming.
- Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)* 71(2), 319–392.
- Scott, S. L., A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch (2016). Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management* 11(2), 78–88.
- Shahbaba, B., S. Lan, W. O. Johnson, and R. M. Neal (2014). Split Hamiltonian Monte Carlo. *Statistics and Computing* 24(3), 339–349.
- Srivastava, S., V. Cevher, Q. Dinh, and D. Dunson (2015). WASP: Scalable Bayes via barycenters of subset posteriors. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 912–920.
- Stan Development Team (2014). Stan: A C++ library for probability and sampling, version 2.5.0.
- Tan, L. S. and D. J. Nott (2013). Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science* 28(2), 168–188.
- van der Geer, S. and J. Lederer (2013). The Bernstein-Olicz norm and deviation inequalities. *Probability letters and related fields* 157, 225–250.
- Wainwright, M. J. and M. I. Jordan (2008, January). Graphical Models, Exponential Families, and Variational Inference. *Found. Trends Mach. Learn.* 1, 1–305.
- Wand, M. (2015). *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*. R package version 2.23-14.
- Wand, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *The Journal of Machine Learning Research* 15(1), 1351–1369.
- Wang, X. and D. B. Dunson (2013). Parallel MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*.
- Wang, X., F. Guo, K. A. Heller, and D. B. Dunson (2015). Parallelizing MCMC with random partition trees. In *Advances in Neural Information Processing Systems*, pp. 451–459.
- Welling, M. and Y. W. Teh (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688.
- Wong, W. H. and X. Shen (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics* 23(2), 339–362.